The Hierarchical Testlet Response Time Model:

Bayesian analysis of a testlet model for item responses and response times

By

Suk Keun Im

Submitted to the graduate degree program in Department of Educational Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

William P. Skorupski, Ed.D.,

Chairperson

_____

Neal Kingston, Ph.D.

_____

Bruce B. Frey, Ph.D.

_____

Vicki Peyton, Ph.D.

_____

Carol M. Woods, Ph.D.

Date Defended: _____

The Dissertation Committee for Suk Keun Im

certifies that this is the approved version of the following dissertation:

The Hierarchical Testlet Response Time Model:

Bayesian analysis of a testlet model for item responses and response times

_____

William P. Skorupski, Ed.D.,

Chairperson

Date approved: _____

Abstract

Computer-based testing makes it possible to record an examinee's response time on an item. This information can be an important factor to understand the examinees, as well as the items (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; van der Linden, 2007). Most response time scoring models are based on unidimensional Item Response Theory (IRT) models. If tests are composed of testlet items, then the assumption of local independence for IRT models is likely to be violated. The purpose of this study is to introduce the Hierarchical Testlet Response Time (HTRT) model to address local dependence among items, and to evaluate the impact on parameter estimation when fitting a response time model to item response and response time data that have been influenced by testlet effects. The study compares the HTRT model with the Hierarchical Framework model (van der Linden, 2007), and explores the relationship between item characteristics and examinee ability as well as response time, which is examined using real and simulated data. The Bayesian estimation using the Markov Chain Monte Carlo (MCMC) method was applied to the investigation of response time. The HTRT model generated better parameter recovery than the Hierarchical Framework model. The HTRT model recovered all parameters very well, with a small magnitude of errors. The current results demonstrate that the Hierarchical Framework model had very good recovery of both the item difficulty and time intensity parameters, but fairly poor recovery of the item discrimination and time discrimination parameters. The examinee ability and speed parameters showed poor recovery, due not to bias but to dramatically increase random error.

Acknowledgements

First and foremost, I would like to thank God for everything. There are several people to whom I am indebted for their contributions. I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my family. My greatest debt is to Dr. William Skorupski, my academic advisor, for his excellent guidance, encouragement, and support throughout my years at the University of Kansas. I also want to thank each member of my dissertation committee, Dr. Neal Kingston, Dr. Bruce Frey, Dr. Vicki Peyton, and Dr. Carol Woods, for support during the course of my study in the program. Additionally, I would like to thank Dr. Marianne Perie, who helped me obtain the dissertation data. Everyone always supported me and encouraged me with their best wishes.

I was able to finish the dissertation with the support and cooperation of my family. I am very thankful to my wife, Bockim Lee, for her loving encouragement and lasting support. I am also blessed with my son and daughter, Andrew and Ashley, for giving me support every day. Finally, I would like to express my deep gratitude to my parents and parents-in-law.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Statement of problem

Traditionally, the goal of testing is to measure how accurately, rather than how quickly, an examinee responds to items. The accuracy of responses has been the main focus in educational assessment (Lee & Chen, 2011). Traditional testing is similar to a power test described by Gulliksen (1950), which measures how accurately an examinee responds to items. However, most educational assessments are neither pure power tests nor pure speed tests because of the involvement of accuracy and time limits.

Computer-based testing makes possible to record an examinee's response time on an item. The response time implies how much time an examinee spends on an item. It was nearly impossible to measure the response time at the individual item level with paper and pencil testing. The additional information of response time can play a significant part in developing, administrating, and validating the test (Zenisky & Baldwin, 2006). This information can be an important factor to understand the examinees, as well as the items (Marianti, Fox, Avetisyan Veldkamp, & Tijmstra, 2014; van der Linden, 2007). Since response time data became available with computer-based testing, there have been numerous research topics involving response time in the field of educational measurement. Some of the topics are aberrant responses (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; van der Linden & van Krimpen-Stoop, 2003), estimating testing time/setting time limits (Bergstrom, Gershon, & Lunz, 1994; Halkitis, Jones, & Pradhan, 1996), examinee motivation/response validity (Wise & DeMars, 2006; Wise & Kong, 2005), item selection in computerized adaptive testing (van der Linden, 2005, 2006), and speededness in computerized adaptive testing (van der Linden, 2009; van der Linden, Scrams, & Schnipke, 1999; van der Linden & Xiong, 2013).

A common topic in educational measurement is scoring models. To obtain valid and accurate estimates of item characteristics and examinee abilities, many models have been introduced. The Item Response Theory (IRT) models have been a popular choice for estimating item characteristics and examinee abilities. The IRT models assume an association between an examinee's responses to items and the underlying latent ability that is measured by the items. When response time data became available, several scoring response time models were developed (e.g., Roskam, 1997; Samejima, 1973, 1974, 1983; Scheiblechner, 1979, 1985; Tatsuoka & Tatsuoka, 1980; Thissen, 1983; van der Linden, 2007; Verhelst, Verstralen, & Jansen, 1997; Wang & Hanson, 2005). The scoring response time models differ by the model's response time distribution, the relationship between examinees' ability and speed, and the model's intended item types. Van der Linden (2009) classified the response time models into three categories. The first category employed distinct models for responses and response times (e.g., Rasch, 1960; Tatsuoka & Tatsuoka, 1980; van der Linden, 2006). The second category used model integration in which response models incorporate response times (e.g., Roskam, 1987, 1997; Verhelst, Verstralen, & Jansen, 1997; Wang & Hanson, 2005). The third category included model integration in which response time models incorporate responses (e.g., Gaviria, 2005; Thissen, 1983). There is no agreement on which model to use with data involving response times. Recently, Suh (2010) claimed that the Hierarchical Framework model (van der Linden, 2007) presents the most reasonable outcomes in both real and simulated data when compared with other response time models. The Hierarchical Framework model was introduced to have response and response time models for each combination of item and person as a first level. The second level includes item domain and population parameters from the two first level models and their relationships.

The key here is that most scoring response time models are based on unidimensional IRT models. Local independence and dimensionality are the important assumptions of unidimensional IRT models. If the unidimensionality assumption is met, then local independence assumption is met because a single latent trait is influencing item responses. The models assume that, given the latent trait parameter, all item responses are statistically independent (Birnbaum, 1968; Lord, 1980). However, there are some circumstances for which the assumption is likely to be violated. One common violation is when tests are composed of testlets (Thissen, Steinberg, & Mooney, 1989; Wainer, Bradlow, & Wang, 2007; Wainer & Lewis, 1990; Wang, Bradlow, & Wainer, 2002). The term 'testlet' was first introduced by Wainer and Kiely (1987) to refer to "a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow" (p. 190). In other words, a testlet is a group of questions linked to a single topic or common stimulus. Within a testlet, responses to items are likely to be dependent on other items, even after controlling for the latent trait. Previous research has clearly established that when tests are constructed with testlet items, local dependence tends to be present among items with a common stimulus (Bradlow, Wainer, & Wang, 1999; Keng, Ho, Chen, & Dodd, 2008; Li, Bolt, & Fu, 2006; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wang, Bradlow, & Wainer, 2002; Wainer & Kiely, 1987). The dependency among testlet items will cause item responses to be more related to each other than a single latent trait alone can explain. The degree of dependence among the items within a testlet depends on the level of variance. A variance of zero indicates that items are locally independent. The greater the variance, the larger the testlet effect.

There have been various suggestions on how to handle violations of the local independence and unidimensionality assumptions. Thissen, Steinberg, and Mooney (1989)

suggested treating testlets as polytomous items, and using polytomous IRT models. However, this approach uses the same discrimination parameter for all items within a testlet and has a total score for each testlet (Zenisky, Hambleton, & Sireci, 2002). These issues may cause a loss of measurement information by having fewer parameters and ignoring the information represented in different scoring patterns. Testlet Response Theory (TRT) models were introduced as a model-based approach to handle the local independence violation (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wang & Wilson, 2005). The traditional unidimensional IRT model is extended to include an additional parameter, γ, to account for within-testlet local dependence. The TRT models are a kind of constrained, confirmatory, multidimensional IRT (MIRT) model, in which all item responses are influenced by a common latent trait, and item responses within a testlet are further explained by a random testlet-effect parameter.

Purpose

The purpose of this study is to introduce the Hierarchical Testlet Response Time (HTRT) model to address the local dependence among items, and to evaluate the impact on parameter estimation when fitting a response time model to data with item responses and response times that have been influenced by testlet effects. The HTRT model is based on the Hierarchical Framework model proposed by van der Linden (2007) and incorporates the approaches of TRT models. The traditional response time models like the Hierarchical Framework model are based on unidimensional IRT model and do not account for the testlet effects. When tests are composed of testlet items, the traditional response time models are likely to violate the assumption of local independence and provide inaccurate parameter estimates. The comparison between the HTRT model and the Hierarchical Framework model was made to explore the relationship among

parameters of responses and response times using real and simulated data. The Bayesian estimation using the Markov Chain Monte Carlo (MCMC) method was applied to investigate and explore the parameters' recovery. This study tried to quantify the estimation errors as a function of various testlet effects, in the context of different test conditions.

Research Questions

The research questions addressed in this study are as follows:

1. If the local independence assumption is violated, then how much improvement does the HTRT model provide over the Hierarchical Framework model (van der Linden, 2007) in parameter estimation?
2. How do various test conditions impact parameter estimation and possibly cause estimation errors?

Hypotheses

The Hierarchical Framework model provided positive results with real data as well as simulated ones (Fox, Klein Entink, & van der Linden, 2007; Suh, 2010; van der Linden, 2007). However, if the local independence assumption is violated or if local dependence among items is present, then the Hierarchical Framework model cannot account for the item dependency. The HTRT model will not be affected by the presence of testlet variances and will account for local dependence among items. The HTRT model will improve the parameter estimation when the local independence assumption is violated.

Chapter 2. Literature Review

Computer-based testing (CBT) makes it possible to record an examinee's response time on an item. This information can be an important factor to understand the examinee, as well as the item (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; van der Linden, 2007). There have been various scoring models to account for response time in recent years (e.g., Thissen, 1983; van der Linden, 2007; Wang & Hanson, 2005). However, these models are based on unidimensional IRT models. This chapter summarizes the relevant topics necessary to understand the intent and outcome of this study. The first section provides an introduction to response time studies. The second section describes a review of IRT models, local dependence, and TRT models. The third section discusses scoring response time models. The final section contains a brief description of Bayesian analysis with the MCMC methods using Gibbs sampling.

Response Time

Response time has been studied in psychology for years (Luce, 1986). The amount of time to respond may provide important information about how an examinee processes information or behaves when responding. Early studies of psychological testing (e.g., Spearman, 1927) assumed that speed and accuracy measure the same construct (Schnipke & Scrams, 2002). Based on this assumption, the scale, speed or accuracy, for measuring one's ability should not matter. In other words, ability should be equivalent whether it is measured on the scale of speed or accuracy.

The goal of testing is usually to measure how accurately an examinee responds to items. Prior to CBT, it was nearly impossible to measure response time to an item. CBT makes it possible to record an examinee's response time on an item. The additional information of

response time can play a significant part in developing, administrating, and validating the test (Zenisky & Baldwin, 2006). This information can be an important factor for understanding the examinees, as well as the items (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; Schnipke & Scrams, 1999; van der Linden, 2007).

For administrative purposes, more power tests are administered with time limits (Morrison, 1960). However, if speed and accuracy are measuring the same construct, then time limits should not affect measurement of ability (Schnipke & Scrams, 2002). Gulliksen (1950) described two different types of test. The pure power test is a test with unlimited time but a fixed number of items with a range of difficulties. The goal is to measure how accurately an examinee responds to items. A power test is similar to traditional testing in that the purpose is usually to measure ability using examinees' responses to items, or accuracy. A pure speed test is a test with an unlimited number of items but a fixed amount of time. The purpose is to measure how quickly an examinee responds to items. In educational assessment, the idea of pure power or pure speed test is very unrealistic (van der Linden & Hambleton, 1997). Recent studies also demonstrated that speed and accuracy do not measure same construct on complex tasks, and that speed does not correlate with score (Baxter, 1941; Bridges, 1985; Foos, 1989; Myers, 1952; Schnipke & Scrams, 2002). Thus, time limits influence examinees' speed and should be considered while measuring one's ability (Bontempo & Julian, 1997).

The speed–accuracy tradeoff is very well known in response time research (Luce, 1986). It implies that accuracy depends on speed (van der Linden, 2009). There is a negative relationship between the speed and accuracy, and one can decide to work at a higher speed with lower accuracy or at a lower speed with higher accuracy (van der Linden, 2007). The speed-accuracy tradeoff is a within-person relationship for cognitive psychologists, while current

studies in psychometrics aim to describe the across-person relationship (Schnipke & Scrams, 2002). The across-person relationship focuses on the relationship between speed and accuracy for a group of examinees. The speed–accuracy tradeoff in testing could be also called the speed–ability tradeoff because ability is the only person parameter to regulate an error probability on test items (van der Linden, 2009; van der Linden, 2011). The trade-off implies that the test taker's ability level during the test depends on his or her choice of speed.

According to Yamamoto and Everson (1997), most standardized educational assessments have time limits. However, a concern of test fairness may be raised from having time limits serving only an administrative purpose (Bridgeman, 2000). Computer adaptive testing (CAT) presents items to examinees that match their ability. In CAT, examinees are unable to omit items. Numerous studies indicated that examinees with higher ability usually take more time to finish items and tests (Bergstrom, Gershon, & Lunz, 1994; Bridgeman & Cline, 2004; Chang, 2006; Swygert, 1998), more difficult items require an additional time for examinees to respond (Bergstrom, Gershon, & Lunz, 1994; Bridgeman & Cline, 2004; Chang, 2006; Plake, 1999; Smith, 2000), and examinees take more time on items to which they respond incorrectly (Bergstrom, Gershon, & Lunz, 1994; Hornke, 2000). In CAT, examinees with higher ability are likely to get more difficult items, and require an extra time to complete the test. Since they are presented with more difficult items, their probability for an incorrect response should be higher. Bridgeman and Cline (2004) concluded that the rapid guessing behavior is more common for higher ability examinees because they receive more difficult items. According to Wise and Kong (2005), rapid-guessing behavior describes examinees rapidly responding to items when they do not have enough time to fully consider the items.

Time limits can influence examinees' scores. Test speededness is when examinees do not have sufficient time to answer all test items within time limits (Bejar, 1985; van der Linden, 2011). Classically, the number of omitted or randomly guessed items at the end of a test were used to measure test speededness. Other methods for measuring test speededness include the ratio of unattempted items at the end to items not answered correctly (Gulliksen, 1950), the correlation between number-correct scores on the same test administered once with time limits and once without time limits (Cronbach & Warrington, 1951), or an amount of time for 80% or 90% of examinees to finish the test (Yamamoto & Everson, 1997). Speededness can impact test scoring, parameter estimation, and test equating (Bejar, 1985; Bridgeman & Cline, 2004; Evans & Reilly, 1972; Kingston & Dorans, 1984; Oshima, 1994; van der Linden, Breithaupt, Chuah, & Zhang, 2007; Wollack, Cohen, & Wells, 2003; Yamamoto & Everson, 1997). However, until the response time from computer-based testing became available, speededness in testing was not taken seriously (Schnipke & Scrams, 2002). Even though most tests contain both power and speed components (Rindler, 1979), test scoring does not usually consider both ability and speed. According to van der Linden (2011), the speed on a test is "the rate of change in the amount of labor performed on the items with respect to time" (p. 46). Ignoring the speed issue can threaten the validity of a test score, and response time models have been developed to accommodate the issue (e.g., Roskam, 1987, 1997; Thissen, 1983; van der Linden, 2007; Verhelst, Verstralen, & Jansen, 1997; Wang & Hanson, 2005). For example, van der Linden's (2007) Hierarchical Framework has separate item and person parameters for both responses and response times to identify differential speededness and control the level of test speededness (van der Linden, 2007; van der Linden, Breithaupt, Chuah, & Zhang, 2007).

Item Response Theory

IRT is a statistical theory that assumes the probability of correct responses is determined by examinees' ability and item characteristics (Birnbaum, 1968; Lord, 1980). IRT creates a scale with useful properties to explain the assessments. The unidimensional models have fundamental assumptions (Birnbaum, 1968; Chen & Thissen, 1997; de Ayala, 2009; Hambleton & Jones, 1993; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). The unidimensionality and local independence assumptions indicate that if test items are measuring a single construct then item response of an examinee are independent of one another, given the examinee's ability. An examinee's ability should be the only factor affecting item responses. The logistic function specifies a monotonically increasing function, such that higher ability results in a higher probability of success. As long as the model fits, item parameters remain unchanged across groups of examinees, and ability parameters remain invariant across groups of items.

There have been numerous IRT models presented since IRT was first introduced. In general, there are two common types of IRT models based on types of item responses: dichotomous or polytomous. The dichotomous IRT models, with binary item responses, include one-, two-, and three-parameter logistic (1-, 2-, and 3-PL) models:

$$\text{1-PL: } P_{ij}\left(u_{ij}=1\middle|\theta_i\right)=\frac{e^{(\theta_i-b_j)}}{1+e^{(\theta_i-b_j)}}$$

$$\text{2-PL: } P_{ij}\left(u_{ij}=1\middle|\theta_i\right)=\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}}$$

$$\text{3-PL: } P_{ij}\left(u_{ij}=1\middle|\theta_i\right)=c_j+(1-c_j)\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}}$$

where $P_{ij}(u_{ij} = 1|\theta_i)$ is the probability that examinee *i* with ability θ_i will have the response $u_{ij} = 1$ to test item *j*. The item parameters include the difficulty or location (b_j), the discrimination or slope (a_j), and the pseudo-guessing or lower asymptote (c_j). The person parameter (θ_i) is interpreted as the ability or proficiency. The 1-PL and 2-PL IRT models are considered constrained forms of 3-PL model. The 2-PL model is equivalent to 3-PL model with all $c_j = 0$. The 1-PL model assumes that the pseudo-guessing parameter is not present and all items have an equal discrimination (i.e., all $c_j = 0$ and all $a_j = 1$).

The polytomous IRT models are applied to tests items with responses of more than two categories. The Graded Response Model (GRM; Samejima, 1969, 1972, 1997) specifies the probability of scoring in a given category or higher given level of ability:

$$P^*_{x_{ij}}(\theta_i) = P(X_{ij} \geq x_{ij}|\theta_i) = \frac{e^{Da_j(\theta_i - b_{x_{ij}})}}{1 + e^{Da_j(\theta_i - b_{x_{ij}})}}$$

where $(x_{ij} = 0, \ldots, M)$, with M being the highest score possible on the item.

The Nominal Response Model (NRM; Bock, 1972) was introduced for item responses in the form of nominal categories. The model specifies the likelihood that an examinee of a given ability will select option k_j of item *j*:

$$P(X_{ij} = k|\theta_i) = \frac{e^{a_{jk}(\theta_i - b_{jk})}}{\sum_{h=1}^{m} e^{a_{jk}(\theta_i - b_{jk})}}$$

The Generalized Partial Credit Model (GPCM; Muraki, 1992, 1993, 1997) is a generalization of the Partial Credit Model (PCM; Masters, 1982; Masters & Wright, 1997), with an item discrimination parameter added to the model. It estimates the probability of getting a score of x rather than $x - 1$:

$$\frac{P_{jx}(\theta)}{P_{j(x-1)}(\theta) + P_{jx}(\theta)} = \frac{e^{a_j(\theta - b_{jx})}}{1 + e^{a_j(\theta - b_{jx})}}$$

When $a_j = 1$ for all n items (i.e., no discrimination parameter), the GPCM is equivalent to the PCM.

Local Dependence

Local dependence is defined as "consistency among item responses that is not accounted for by individual differences on the construct we intend to measure" (Steinberg & Thissen, 1996, p. 83). Yen (1993) described possible causes for local dependence (e.g., speededness, fatigue, passage dependence, scoring rubric, and practice). Thissen, Steinberg, and Mooney (1989), Wainer (1994), and Wainer and Lewis (1990) assumed that a passage or common stimulus followed by a number of items may violate the assumption of conditional independence. Hosken and De Boeck (1997) demonstrated two types of local dependence. "Order dependency" is when item responses are influenced by earlier item responses. "Combination dependency" is when a common stimulus like paragraph or graph is followed by a set of items.

There have been several suggestions on how to measure local dependence. Yen (1984) proposed Q_3 statistics, which indicate the correlation between two items after accounting for overall performance. Chen and Thissen (1997) proposed G^2 LD index, which analyzes the residuals from the IRT model for each pair of items. Steinberg and Thissen (1996) mentioned that factor analysis could be a possible way to detect local dependence, but items may have many different sources of local dependence.

There are various suggestions on how to deal with local dependence. For reading comprehension tests, a testlet (instead of an item) should be considered the unit of analysis

(Sireci, Thissen, & Wainer, 1991). Yen (1993) recommended creating independent items, combining locally dependent items for grading, constructing separate scales, or using teslets.

In the presence of local independence violations, using traditional IRT models will produce flawed parameter estimation (Chen & Wang, 2007; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Tuerlinckx & De Boeck, 2001; Wainer & Thissen, 1996; Wang, Baldwin, Wainer, Bradlow, Reeve, Smith, Bellizzi, & Baumgarter, 2010; Yen, 1993). Ignoring local dependence can lead to overestimates of reliability, underestimates of standard error of ability estimates, and underestimates of item discrimination (Bradlow, Wainer, & Wang, 1999; Sireci, Thissen, & Wainer, 1991; Wainer & Wang, 2000; Yen, 1993). Wainer and Wang (2000) indicated that item difficulties were well estimated but lower proficiency levels were overestimated when the local independence assumption was violated. Wainer, Bradlow, and Du (2000) revealed that the difficulties and examinee abilities were recovered better than the item discriminations and lower asymptotes with locally dependent items. However, Bradlow, Wainer, and Wang (1999) and Wainer, Bradlow, and Wang (2007, p. 106) mentioned that, even though the number of items per testlet may not correlate with incorrect estimation, the amount of error can be minimized with a testlet size of 4 to 6 items per testlet.

Testlet Response Theory

The local independence assumption is one of the most important assumptions in IRT modeling. However, there are some circumstances for which the assumption of local independence is likely to be violated. One common violation is when tests are composed of testlets (Wainer, Bradlow, & Wang, 2007; Wang, Bradlow, & Wainer, 2002). Wainer and Kiely (1987) introduced the term ‘testlet’ to refer to “a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee

may follow" (p.190). Typically, a testlet is a group of items related to a single common stimulus. Based on the mathematical definition of a testlet presented in TRT (Wainer, Bradlow, & Wang, 2007), a testlet can be as small as a single item or as large as the entire test, though typically testlets are subsets of items within a test.

When tests are constructs with testlet items, local dependence is likely to appear among testlet items (Bradlow, Wainer, & Wang, 1999; Keng, Ho, Chen, & Dodd, 2008; Li, Bolt, & Fu, 2006; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wang, Bradlow, & Wainer, 2002; Wainer & Kiely, 1987). If testlet effects are ignored, the violation of local independence with unidimensional IRT models may lead to incorrect parameter estimation (Sireci, Thissen, & Wainer, 1991; Wainer & Thissen, 1996; Yen, 1993). However, Wainer, Bradlow, and Wang (2007) suggest that traditional IRT models where the units of measure are testlets can account for the violation of local independence.

There have been various suggestions on how to handle testlet effects. Thissen, Steinberg, and Mooney (1989) suggested treating testlets as polytomous items and applying polytomous IRT models. However, this approach uses the same discrimination parameter for all items within a testlet and a total score for each testlet (Zenisky, Hambleton, & Sireci, 2002). These issues will possibly cause a loss of measurement information by having a fewer parameters, and different scoring patterns for each testlet will be ignored.

Bradlow, Wainer, and Wang (1999) introduced the TRT as 2-PL model for dichotomous items to handle the local independence violations. The TRT model is a kind of constrained, confirmatory, multidimensional IRT (MIRT) model, in which all item responses are influenced by a common latent trait, and item responses within a testlet are further explained by a random testlet-effect parameter. Wang and Wilson (2005) introduced the Rasch Testlet Model for both

dichotomous and polytomous items, noting that it is a special case of the multidimensional random coefficients multinomial logit model (MRCMLM; Adams, Wilson, & Wang, 1997). Wainer, Bradlow, and Du (2000) introduced the 3-PL TRT model, which is a simple extension of the standard 3-PL IRT model. Wang, Bradlow, and Wainer (2002) further extended the TRT model to include mixed format assessments, allowing both dichotomous and polytomous responses.

For all approaches to the TRT models, the unidimensional IRT models are extended to include an additional parameter, γ, to account for within-testlet local dependence. The 3-PL TRT model is:

$$P_{ij}\left(u_{ij}=1\middle|\theta_i\right)=c_j+(1-c_j)\frac{e^{a_j\left(\theta_i-b_j+\gamma_{id(j)}\right)}}{1+e^{a_j\left(\theta_i-b_j+\gamma_{id(j)}\right)}}$$

where $P_{ij}(u_{ij}=1|\theta_i)$ is the probability of an examinee with the proficiency θ_i having a response $u_{ij}=1$ to item j. The interpretation of the *a*-, *b*-, and *c*-parameters is the same as for the 3-PL IRT model. The testlet effect parameter, $\gamma_{id(j)}$, represents the interaction between an examinee *i* and items within testlet *d(j)*, where *d* is a vector of categorical integers indicating to which testlet each item belongs. The addition of the $\gamma_{id(j)}$ parameter to the model allows items within a testlet to have a higher marginal correlation, which represents the interaction between an examinee *i* and items within a testlet. As such, the TRT models are special cases of MIRT models, allowing for a confirmatory approach to modeling multiple "abilities" (θ and γ) for each examinee, but constraining the model to have a single discrimination parameter per item. In order to identify the model, $\gamma_{id(j)}$ is constrained to have a mean of zero within testlets. The influence of the testlet parameter is evidenced by their within-testlet variances. The larger the variance, the larger the testlet effect is. If the traditional identification constraint is made on θ_i,

(mean=0, variance=1), then the variance of $\gamma_{id(j)}$ can be interpreted on that metric as how much additional dimensionality is explained by the testlet effect.

Response Time Models

With more tests administered on computers, it has become easier to collect response times on each item. Recent response time models focus more on empirical response time distributions. Scrams and Schnipke (1997) suggested the comparison of speed and accuracy as separate components of proficiency by using response times in standardized tests. Instead of using only item responses, response time models are using both item responses and response times to measure ability. Traditional IRT models were adjusted to incorporate response times.

Van der Linden (2009) reviewed different categories of response time modeling. The first category is distinct models for response times and responses (e.g., Rasch, 1960; Tatsuoka & Tatsuoka, 1980; van der Linden, 2006). The second category is model integration in which response models incorporate response times (e.g., Roskam, 1987, 1997; Verhelst, Verstralen, & Jansen, 1997; Wang & Hanson, 2005). As an extension of the second category, the third category is model integration in which response time models incorporate responses (e.g., Gaviria, 2005; Thissen, 1983).

The Hierarchical Framework was introduced to by van der Linden (2007) to have response and response time models for each combination of person and item as a first level, and population and item domain parameters from the two first level models and their relationships as a second level. Figure 1 shows a graphical representation of the Hierarchical Framework model. Suh (2010) demonstrated that the Hierarchical Framework model presents the most reasonable outcomes in both real data and simulated when compared with other popular response time models.

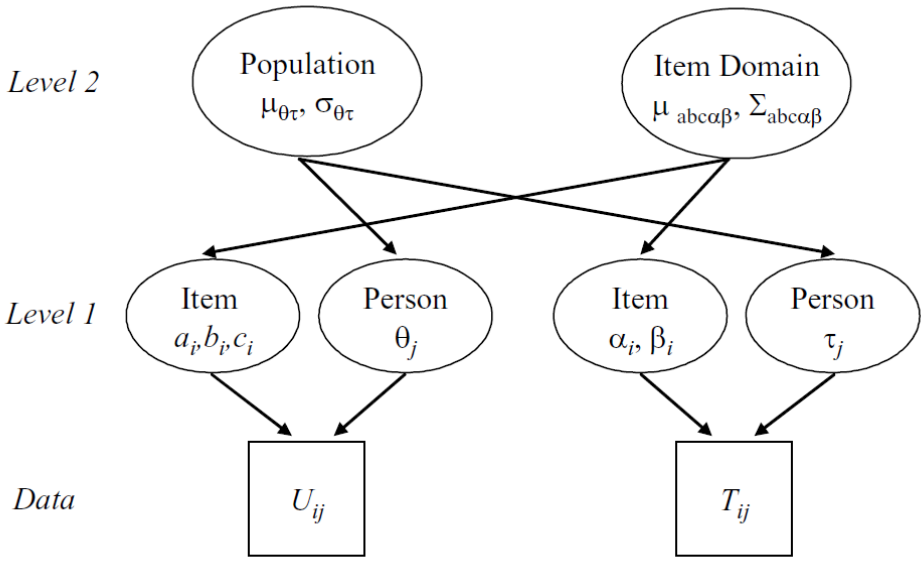

Figure 1. *The hierarchical framework for modeling speed and accuracy on items (van der Linden, 2007).*

There are important assumptions for the Hierarchical Framework model (van der Linden, 2007; van der Linden, 2009). These assumptions include that examinees take the test at a fixed speed level, that item response and item response time are considered to be random variables, that there are separate item and person parameters for both the response and response time models, that responses and response times are conditionally independence given the levels of ability and speed, and that the relationship between speed and accuracy will be modeled for both the population of examinees and a single examinee.

In the first level, the response model is the 3-PL IRT model, with the usual parameters and interpretations.

$$P(u_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}$$

where $u_{ij} = 1$ is an examinee *i* having a correct response to item *j*, θ_i is the ability parameter for an examinee *i*, a_j is the discrimination parameter for item *j*, b_j is the difficulty parameter for item *j*, and c_j is the pseudo-guessing parameter for item *j*.

A lognormal model is chosen for the response times:

$$f(t_{ij}; \tau_i, \alpha_j, \beta_j) = \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_j\left(ln\, t_{ij} - (\beta_j - \tau_i)\right)\right]^2\right\}$$

where t_{ij} is a response time by examinee *i* on item *j*, τ_i is the speed parameter of examinee *i*, β_j represents the time intensity of item *j*, and α_j represents the discriminating power of item *j*. The larger the speed parameter, the faster an examinee is operating during the test. Larger time intensity parameters indicate larger amounts of time examinees spend on the item. Since response time has a natural lower bound of zero, the guessing parameter is not needed.

In the second level, the joint distribution of the person parameters are referred to as the population model:

$$\xi_i \sim f(\xi_i; \mu_p, \Sigma_p)$$

where ξ_i is the vector of person parameters θ_i and τ_i, and assumed to have multivariate normal distribution with

$$\mu_p = (\mu_\theta, \mu_\tau)$$

$$\Sigma_p = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}$$

The joint distribution of the item parameters are stated as the item-domain model:

$$\psi_j \sim f(\psi_j; \mu_I, \Sigma_I)$$

where ψ_j is the vector of item parameters $a_j, b_j, c_j, \alpha_j,$ and β_j, and assumed to have multivariate normal distribution.

$$\mu_I = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta)$$

$$\Sigma_I = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\beta\alpha} & \sigma_\beta^2 \end{pmatrix}$$

For identification, van der Linden (2007) suggested the following constraints.

$$\mu_\theta = 0, \sigma_\theta^2 = 1, \mu_\tau = 0$$

The first two constraints are common in IRT models. The third constraint removes the tradeoff between β_j and τ_i from a lognormal model in the first level. These constraints allow all covariances between item and person parameters to be freely estimated (van der Linden, 2007). The scale of the response time parameter is fixed by the time unit, $\ln t_{ij}$.

Van der Linden (2007) applied 3-PL IRT model as the response model; however, it can be replaced by any other IRT scoring models (e.g., 1-PL, 2-PL, polytomous, or multidimensional). Because the response and response time models in the first level are separate, the response model can be replaced without changing the response time model.

For the population and item-domain models, normal/inverse-Wishart prior distributions were chosen:

$$\Sigma_P \sim \text{Inverse-Wishart } (\Sigma_{P0}^{-1}, \upsilon_{P0})$$

$$\mu_P | \Sigma_P \sim \text{MVN}(\mu_{p0}, \Sigma_P / \kappa_{P0})$$

$$\Sigma_I \sim \text{Inverse-Wishart } (\Sigma_{I0}^{-1}, \upsilon_{I0})$$

$$\mu_I | \Sigma_I \sim \text{MVN } (\mu_{I0}, \Sigma_I / \kappa_{I0})$$

where υ_{P0} and υ_{I0} are scalar degrees of freedom parameters, Σ_{P0} and Σ_{I0} are scale matrices for the prior on Σ_P and Σ_I, μ_{P0} and μ_{I0} are vectors with means of posterior distributions, and κ_{P0} and κ_{I0} are the strength of prior information about these means.

A common prior distribution for the guessing parameters in the first-level model was used:

$$c_j \sim \text{beta}\ (\gamma, \delta)$$

According to Patz & Junker (1999a), it is possible for MCMC method to have difficulty with the weak identifiability of the 3-PL model. For the potential difficulty, van der Linden (2007) fixed $c_j = 0.20$ to address a possible tradeoff between a_j and c_j parameters.

Bayesian Analysis

Both the TRT and the Hierarchical Framework models are developed and embedded in the Bayesian context (Bradlow, Wainer, & Wang, 1999; van der Linden, 2007). The goal of Bayesian analysis is to fit a probability model to data, and summarize the results by a probability distribution of parameters (Fox, 2010; Gelman, Carlin, Stern, & Rubin, 2003; Lynch, 2007). Bayesian analysis makes probability inferences about some unobserved parameter θ conditional on observed data y, that is, $P(\theta|y)$. Generally, Bayesian analysis is conducted with three steps. First, a full probability model is set up to specify the joint probability distribution for all observable and unobservable quantities. Second, a posterior distribution is estimated by conditioning on observed data. Finally, the assumptions and fit are evaluated.

The joint probability distribution of θ and y is the product of the prior distribution $P(\theta)$ and the sampling distribution (or likelihood function) $P(y|\ \theta)$:

$$P(\theta, y) = P(\theta)P(y|\theta)$$

The posterior distribution is produced by applying Bayes' rule, conditioning on observed data y to calculate a probability distribution (or posterior distribution) for the unobserved parameter:

$$P(\theta|y) = \frac{P(\theta, y)}{P(y)} = \frac{P(y|\theta)P(\theta)}{P(y)}$$

$P(y)$ is the marginal distribution of y, obtained by integrating over $P(\theta, y)$ with respect to θ.

$$P(y) = \Sigma_\theta P(y|\theta)P(\theta) \text{ for discrete } \theta$$

$$P(y) = \int_\theta P(y|\theta)P(\theta)d\theta \text{ for continuous } \theta$$

With fixed y, $P(y)$ does not depend on θ and becomes a constant, and can be removed:

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

The prior distribution $P(\theta)$ represents prior belief or information about the distribution of θ, without considering data. It can be estimated by specifying the density function that θ should follow. Once the prior is specified, the posterior distribution of parameters can be estimated. The prior distributions can be conjugate or non-conjugate to the posterior (Fox, 2010). A conjugate prior has the same parametric form as the posterior and is very computationally convenient. However, in multidimensional examples, conjugacy may not be possible. This leads to the prior having different parametric form than the posterior, and the prior becomes non-conjugate.

Markov Chain Monte Carlo Estimation

According to Gelman, Carlin, Stern, and Rubin (2003), Markov Chain Monte Carlo (MCMC) is "a general method based on drawing values of θ from approximate distributions and then correcting those draws to better approximate the target posterior distribution, $P(\theta|y)$" (pp.

285–286). The concept of MCMC estimation is to construct a set of random draws from the posterior distribution for each parameter being estimated. The random draws are simulated from any theoretical distributions and the features of the theoretical distributions are based on the samples (Patz & Junker, 1999a, 1999b). These draws are either accepted or rejected as being reasonable from the actual posterior distribution until enough draws are retained to make inferences. The MCMC method is easy to implement and free software is available, but generally takes a long time and requires sophisticated algorithms (Kim & Bolt, 2007; Patz & Junker, 1999a, 1999b). It represents an estimation strategy in a perspective of Bayesian inference (Kim & Bolt, 2007). The MCMC method has expanded the opportunity to experiment with new models needed for specialized measurement applications (Kim & Bolt, 2007; Lynch, 2007).

Patz and Junker (1999a) introduced the use of the MCMC method for the IRT models. Kim and Bolt (2007) explained how to implement the MCMC method with IRT models. The priors are necessary for all parameters in the IRT models using the MCMC method. After the IRT model is chosen and priors for all parameters are specified, sampling procedures can be performed.

By applying the MCMC method, some considerations are needed. Those considerations include the number of Markov chains, the length of each chain, the "burn-in" period, and thinning. At least two independent chains are recommended for evidence of convergence, and each chain should be long enough to converge at a stationary distribution. The first K draws are usually discarded as "burn-in" because of unstableness of the Markov Chain at an early stage. The thinning can be considered by keeping every kth simulated draw from each chain. Traditionally, a thin value was considered, but Link and Eaton (2012) claimed that thinning does not provide much improvement in inference-making.

The Markov chain for a given parameter has converged if multiple chains arrive at the same stationary distribution. Once convergence is assumed, the samples from the posterior distribution are used to estimate model parameters. According to Kim and Bolt (2007), there are several factors affecting the convergence rate. The high correlation between adjacent states can cause slow convergence and require very long iterations, and the sampling algorithm and identification can also affect model convergence (Kim & Bolt, 2007; Lynch, 2007).

There are several suggestions on how to detect convergence. First, one can inspect the history of the chain. Then, one can apply a number of convergence diagnostics (e.g., Geweke's (1992) criterion, Raftery & Lewis's (1992) criterion, and Gelman & Rubin's (1992) criterion). Geweke's (1992) criterion computes a z-score from the sampled states for each parameter, with a z-score within the non-significance range taken as evidence of convergence. Raftery and Lewis' (1992) criterion returns an index from considering the number of samples needed to estimate accurate quantiles of the posterior. Index values greater than 5.0 indicate that more sampled states are needed to reach convergence due to autocorrelations in the chain. Gelman and Rubin's (1992) criterion, $\sqrt{\hat{R}}$, can be considered when multiple chains are applied. It compares the variances within and between chains for each parameter, with values close to 1 considered indicative of convergence.

Gibbs Sampler

The Gibbs sampler in Bayesian statistics was introduced by Gelfand and Smith (1990). It is a special case of the more general Metropolis-Hastings algorithm (Lynch, 2007). According to Kim & Bolt (2007), "the Gibbs sampler provides a mechanism by which sampling can be performed with respect to smaller numbers of parameters, often one at a time" (p. 41). Gibbs

sampling is also called alternating conditional sampling (Gelman, Carlin, Stern, & Rubin, 2003, p. 41).

The parameter θ is divided into *d* components, $\theta = (\theta_1, \ldots, \theta_d)$, and each iteration cycles through all *d* components. There are thus *d* steps in iteration *t*. At each iteration *t*, an ordering of the *d* component is chosen and each estimand is sampled from the conditional distribution, given all the other components of θ:

$$P(\theta_j | \theta_{-j}^{t-1}, y)$$

where θ_{-j}^{t-1} represents all the components of θ except for θ_j at their current values:

$$\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1})$$

Thus, each subvector θ_j is conditionally updated on the latest values of the other components of θ, which are the iteration *t* values for the components already updated and the iteration *t*-1 values for the others. The j^{th} estimand at time *t* is conditional on all sampled estimands at time *t* and those that haven't been updated yet, at time *t*-1.

Convergence

$\hat{R}$ (Brooks & Gelman, 1998; Gelman & Rubin, 1992) is also known as a potential scale reduction factor.

$$\sqrt{\hat{R}} = \frac{\sqrt{\hat{V}^+(\psi|y)}}{W}$$

The numerator of the above formula is the marginal posterior variance of each estimated parameter,

$$\hat{V}^+(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

where *n* = chain length, *W* = within-chain variance, and *B* = between-chain variance.

The denominator of the above formula is the within-chain variance. The $\sqrt{\hat{R}}$ value less than 1.2 is considered to be a good evidence of convergence (Brooks & Gelman, 1998; Gelman, 1996, p. 170). If $\sqrt{\hat{R}}$ is greater than 1.2, then a longer length of the Markov Chain should be considered for the convergence.

Deviance Information Criterion

The DIC can be calculated as follows:

$$DIC = \bar{D} + P_D = D(\bar{\theta}) + 2P_D$$

where $D(\bar{\theta}) = -2\ln(L)$ is the posterior expectation of deviance and P_D is the effective number of parameters. The model with a smaller value of DIC indicates to be the better model for the observed data. Estimation of the DIC index can be obtained from the *R2OpenBUGS* package (Sturtz, Ligges, & Gelman, 2010).

Chapter 3. Methods

This study introduces a new response time model, the HTRT model, to address testlet effects. Im and Skorupski (2014) have shown that the amount of parameter estimation error increases with the presence of testlet variance. Researchers are more likely to obtain inaccurate parameter estimates if testlet effects are ignored (Sireci, Thissen, & Wainer, 1991; Wainer & Thissen, 1996; Yen, 1993). The traditional response time models, like the Hierarchical Framework model, do not account for these effects. This study tries to explain the dependence among items using real and simulated data. In Study 1, the HTRT model was applied to real data through Bayesian estimation using the MCMC method. In Study 2, the HTRT model and the Hierarchical Framework model were applied to simulated data. The models were explored using simulated data with known parameters to understand how the models behave under different test conditions. Since Suh (2010) demonstrated that the Hierarchical Framework model presents the most reasonable outcomes in both real and simulated data when compared with other response time models, the HTRT model was based on the response time model by van der Linden (2007).

Study 1

Data

Real data from a reading assessment administered in 2012 from a midwestern U.S. state was used for the study. Data contain both response and response time from operational items. For this study, three grades were selected that contained the most examinees, and one test form was selected per grade. To reduce any missing information affecting the results, this study included only examinees who completed all of the items within test forms. Table 1 presents test

length, sample size, the number of testlets, the number of items per testlet, the number of testlet items, and the number of independent items for each of three grades selected for the study.

Table 1
Information of three grades selected for the study

Grade	Test length	Sample size	Number of testlets	Number of items per testlet	Number of testlet items	Number of independent items
3	54	1,378	6	14, 7, 8, 7, 9, 9	54	0
4	68	1,005	7	12, 8, 11, 7, 7, 11, 8	64	4
5	68	888	8	7, 10, 8, 10, 6, 6, 10, 6	63	5

Examinees' responses are coded dichotomously and response times are recorded in seconds. The recorded response time for an item is the total time spent on the item during all attempts at that item. Since the natural log of zero is undefined, response times of zero were coded as missing. Unrealistic response times for a single item (larger than 1,000,000,000 seconds) were also coded as missing. Personal identifiers such as names, identification numbers, and school information were removed from the data to protect the anonymity of examinees.

Estimation Methods

Figure 2 shows a graphical representation of the HTRT model. The HTRT model is based on van der Linden's (2007) Hierarchical Framework model. The 2-PL Testlet Response Theory (TRT) model is selected as the response model and a lognormal model with an added testlet parameter is selected as the response time model. The studies regarding the Hierarchical Framework model (Fox, Klein Entink, & van der Linden, 2007; Suh, 2010; van der Linden, 2007) used the 2-PL IRT model as their default model. According to Patz and Junker (1999a), the MCMC method may cause difficulty due to weak identifiability with the 3-PL model. For these reasons, the 2-PL TRT model was selected as the default model for the study.

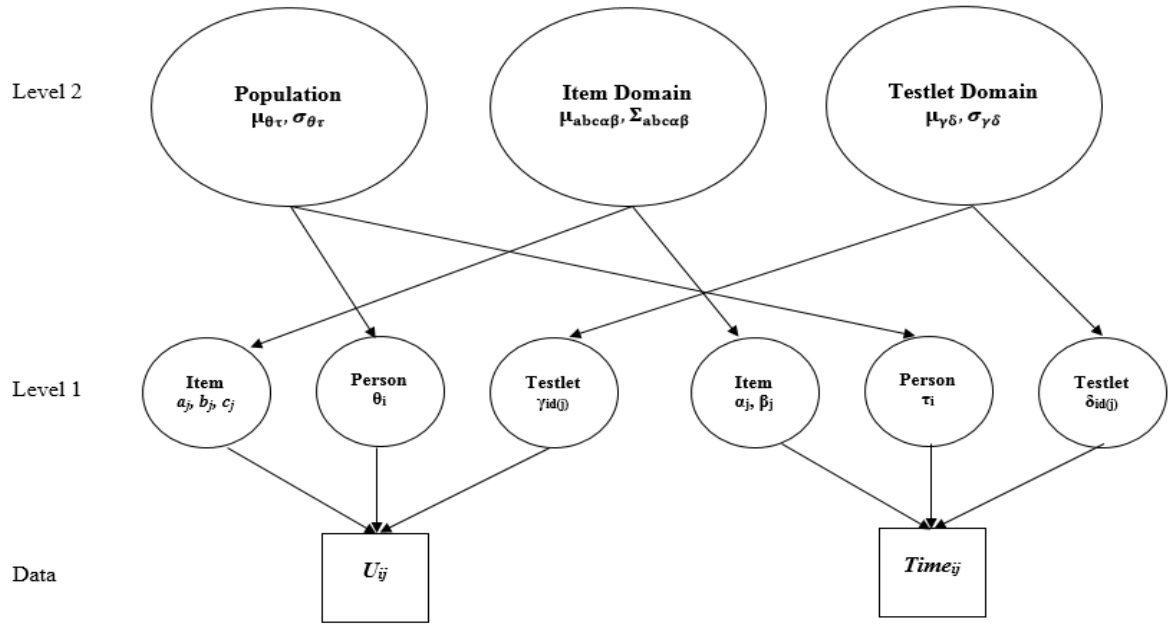

Figure 2. *The graphical representation of the HTRT model.*

The first level includes the TRT model as the response model,

$$P\left(u_{ij}=1\middle|\theta_i,a_j,b_j,\gamma_{id(j)}\right)=\frac{e^{a_j\left(\theta_i-b_j+\gamma_{id(j)}\right)}}{1+e^{a_j\left(\theta_i-b_j+\gamma_{id(j)}\right)}}$$

where $P_{ij}(u_{ij}=1|\theta_i)$ is the probability of an examinee with the proficiency θ_i having a response $u_{ij}=1$ to item j. The item parameters include the difficulty or location (b_j) and the discrimination or slope (a_j). The testlet effect parameter, $\gamma_{id(j)}$, represents the interaction between an examinee *i* and items within testlet *d(j)*, where *d* is a vector of categorical integers indicating to which testlet each item belongs. The person parameter (θ_i) is interpreted as the ability or proficiency.

A lognormal model with a testlet variable is the response time model:

$$f\left(t_{ij};\tau_i,\alpha_j,\beta_j,\delta_{id(j)}\right)=\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_j\left(ln\ t_{ij}-\left(\beta_j-\tau_i+\delta_{id(j)}\right)\right)\right]^2\right\}$$

where t_{ij} is a response time by examinee *i* on item *j*, τ_i is the speed parameter of examinee *i*, β_j represent the time intensity of item *j*, α_j represents the discriminating power of item *j*. The testlet effect parameter, $\delta_{id(j)}$, represents the interaction between an examinee *i* and items within testlet *d(j)*, where *d* is a vector of categorical integers indicating to which testlet each item belongs.

The model is investigated through the Bayesian estimation using the MCMC method. In the Bayesian estimation, prior distributions are specified for model parameters. The parameter priors are set to be fairly large and less informative; thus, the data can drive the posterior distributions. The starting values for each parameter of the Markov Chain are randomly generated using OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009). The estimation was conducted using the *R2OpenBUGS* package (Sturtz, Ligges, & Gelman, 2010) for R (R Core Team, 2014).

This study adopted and made changes to the priors used by van der Linden (2007). The HTRT model is similar to the Hierarhical Framework model but includes γ and δ parameters to represent testlet effects in response and response time models, respectively. The person parameters (θ & τ) and item parameters (a, b, α, and β) used the following priors:

$$\mu_P = (\mu_\theta, \mu_\tau) = (0, 0)$$

$$\Sigma_{P0}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu_I = \left(\mu_a, \mu_b, \mu_\alpha, \mu_\beta\right) = (1, 0, 1, 0)$$

$$\Sigma_{I0}^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

For the testlet parameters (γ & δ) of this study, the following priors were considered:

$$\mu_{Tk} = \left(\mu_{\gamma k}, \mu_{\delta k}\right) = (0, 0)$$

$$\Sigma_T = \begin{pmatrix} \sigma_\gamma^2 & \sigma_{\gamma\delta} \\ \sigma_{\gamma\delta} & \sigma_\delta^2 \end{pmatrix}$$

$$\sigma_\gamma^2 \sim U(0, 10)$$

$$\sigma_\delta^2 \sim U(0, 10)$$

$$\gamma_k \sim N\left(0, \frac{1}{\sigma_\gamma^2}\right)$$

$$\delta_k \sim N\left(0, \frac{1}{\sigma_\delta^2}\right)$$

where k=1,…,K and K = the total number of testlets.

Analysis and Model Convergence

This study set the length of the Markov Chain as 10,000 and increased the length as needed. The length of the Markov Chain needs to be long enough to reach the convergence. The burn-in period and post burn-in period for parameter estimates were determined by visually inspecting the history of parameters. Two independent MCMC chains were run for each parameter, and thinning was not used. According to Link and Eaton (2012), thinning is often unnecessary and inefficient.

To assume draws come from the posterior distribution, the Markov Chains should converge on a stationary distribution. This study used two independent chains for each monitored parameter to check the convergence. In addition to the visual inspections of history plots, autocorrelations, and posterior distributions of parameter estimates, this study used $\sqrt{\hat{R}}$ (Brooks & Gelman, 1998; Gelman & Rubin, 1992) to determine the convergence.

Study 2

The goal of Study 2 is to quantify the estimation error as a function of various testlet effects in the context of different test conditions. A Monte Carlo simulation study was conducted to evaluate the effects of testlet variance levels and different response time models on estimation error. The local independence assumption was intentionally violated by calibrating the simulated data with the Hierarchical Framework model. The data were also calibrated with the HTRT model for the comparison. Simulated data were generated from information collected from Study 1 (i.e., estimated parameters, test formats, test length, and number of examinees).

Design

In order to emulate a testing situation from real data, information from Study 1 was used for Study 2. The estimated parameters, test formats, test length, and number of examinees from Study 1 for each grade were fixed variables for Study 2. The estimated parameters from Study 1 were considered the true parameters.

For each grade, a Monte Carlo simulation study was conducted to evaluate the effects of (1) size of testlet variance on parameter estimation error, and (2) different response time models. Table 2 presents independent variables and their levels.

Table 2
Independent variables and their levels

Model	Variance of γ	Variance of δ
Hierarchical Framework Model	0.25	0.25
		0.50
		1.00
	0.50	0.25
		0.50
		1.00
	1.00	0.25
		0.50
		1.00
Hierarchical Testlet Response Time Model	0.25	0.25
		0.50
		1.00
	0.50	0.25
		0.50
		1.00
	1.00	0.25
		0.50
		1.00

Three testlet variance conditions in the response model, $\sigma_\gamma^2 = (0.25, 0.5, 1.0)$, and three testlet variance conditions in the response time model, $\sigma_\delta^2 = (0.25, 0.5, 1.0)$, were crossed to produce nine testlet variance conditions. These values were selected based on previous research (Bradlow, Wainer, & Wang, 1999; Im & Skorupski, 2014; Wang, Bradlow, & Wainer, 2002) to represent a range of small to large testlet effects.

Data Generation

For each grade, item response data and response time data were simulated from the HTRT model. Every dataset was independently replicated 50 times, and results of the replications were averaged to evaluate the stability of findings. Item responses and response times were generated in *R* (R Core Team, 2014). For every condition and replication, a fixed sample of item, ability, and testlet parameters were used, with new data for every replication.

Analyses

The simulated data were calibrated with the 2-PL Hierarchical Framework model and the 2-PL HTRT model using OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009) with the *R2OpenBUGS* package (Sturtz, Ligges, & Gelman, 2010) for *R* (R Core Team, 2014).

In this study, the person parameters (θ & τ) and item parameters (a, b, α, & β) used the following priors:

$$\mu_P = (\mu_\theta, \mu_\tau) = (0, 0)$$

$$\Sigma_{P0}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu_I = (\mu_a, \mu_b, \mu_\alpha, \mu_\beta) = (1, 0, 1, 0)$$

$$\Sigma_{I0}^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

For the testlet parameters (γ & δ) of the HTRT model, the following priors were selected:

$$\mu_{Tk} = (\mu_{\gamma k}, \mu_{\delta k}) = (0, 0)$$

$$\Sigma_T = \begin{pmatrix} \sigma_\gamma^2 & \sigma_{\gamma\delta} \\ \sigma_{\gamma\delta} & \sigma_\delta^2 \end{pmatrix}$$

$$\sigma_\gamma^2 \sim U(0, 10)$$

$$\sigma_\delta^2 \sim U(0, 10)$$

$$\gamma_k \sim N\left(0, \frac{1}{\sigma_\gamma^2}\right)$$

$$\delta_k \sim N\left(0, \frac{1}{\sigma_\delta^2}\right)$$

where k=1,…,K and K = the total number of testlet.

Measured Criteria

Both models were evaluated by how well they recovered the known parameters through descriptive statistics, bias, mean squared error (MSE), and test information function (TIF). The means and standard deviations of these error indices were computed across 50 replications. For the item and ability parameters, bias was calculated as the mean difference between estimated and true parameters. The MSE was calculated as bias squared and then averaged. The average bias and MSE, which, respectively, represent the systematic and total error variance, were calculated over replications.

Below is the bias formula for the ability (θ) parameter. This formula can be extended to the remaining parameters by substituting each of the other parameters for θ.

$$\boldsymbol{BIAS}_{\theta} = \frac{\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \sum_{\boldsymbol{j}=1}^{\boldsymbol{R}} \left(\hat{\theta}_{\boldsymbol{ij}} - \theta_{\boldsymbol{i}}\right)}{\boldsymbol{NR}}$$

The formula below for the MSE can likewise be altered to establish the MSE for the other parameters.

$$\boldsymbol{MSE}_{\theta} = \frac{\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \sum_{\boldsymbol{j}=1}^{\boldsymbol{R}} \left(\hat{\theta}_{\boldsymbol{ij}} - \theta_{\boldsymbol{i}}\right)^2}{\boldsymbol{NR}}$$

As an overall measure of test structure recovery, the relationship between the true test information function with true parameter estimates and its estimate over replications were also evaluated graphically.

Checking Model Convergence and Model Fit

For the MCMC estimation, the number of chains, number of iterations, and burn-in period were adopted from Study 1. As in Study 1, the MCMC chains in Study 2 were not thinned.

Even though the MCMC components are adopted from Study 1, the $\sqrt{\hat{R}}$ was evaluated for the model convergence. The Deviance Information Criterion (DIC) values from both models were used for the model comparison.

Chapter 4. Results

In this chapter, the results from the real and simulated data are presented. In Study 1, the HTRT model was applied to real data. The overall descriptions of data are presented, and the information about the MCMC estimation components is explored for the subsequent study. For Study 2, the HTRT model and the Hierarchical Framework model were applied to simulated data using the MCMC estimation. The convergence and parameter recovery are examined for various conditions of testlet parameters to compare the two models.

Study 1

Preliminary Data Analysis

Table 3 shows the descriptive statistics for responses and response times, and Figures 3 – 5 show the distributions of total scores and total response times. Table 4 categorizes the number of examinees by total testing time. For all three grade levels, total scores (i.e., sum of correct items) were negatively skewed and total response times (i.e., sum of response times for all items) were positively skewed. All three grades had high average total scores. This may result from the fact that the study included only examinees who responded to all items. Because of the natural lower bound at zero, the distribution of response time is likely to be positively skewed. The average amount of testing time for each grade ranged from 39 to 48 minutes. Most examinees took the assessment within one to two hours. However, there were outliers, with some examinees who took nearly eight hours for the assessment.

Table 3

Descriptive statistics for responses and response times

Grade	Variable	N	Items	Mean	SD	Min	Max	Skewness
3	Total Score	1,378	54	46.30	6.15	14.00	54.00	-1.38
	Total Time	1,378	54	2310.59	2303.69	169.00	29255.00	4.65
4	Total Score	1,005	68	59.69	7.39	24.00	68.00	-1.55
	Total Time	1,005	68	2845.31	2484.20	299.00	19122.00	3.14
5	Total Score	888	68	59.18	5.66	34.00	68.00	-1.37
	Total Time	888	68	2830.19	2262.37	415.00	31764.00	4.28

Table 4

Number of examinees on total testing time in hours

Grade	Hour	Frequency	Percent
3	1	1,161	84.25%
	2	178	12.92%
	3	21	1.52%
	4	6	0.44%
	5	10	0.73%
	6	0	0.00%
	7	0	0.00%
	8	2	0.15%
4	1	795	79.10%
	2	163	16.22%
	3	20	1.99%
	4	15	1.49%
	5	10	1.00%
	6	2	0.20%
5	1	672	75.68%
	2	181	20.38%
	3	26	2.93%
	4	6	0.68%
	5	1	0.11%
	6	1	0.11%
	7	0	0.00%
	8	1	0.11%

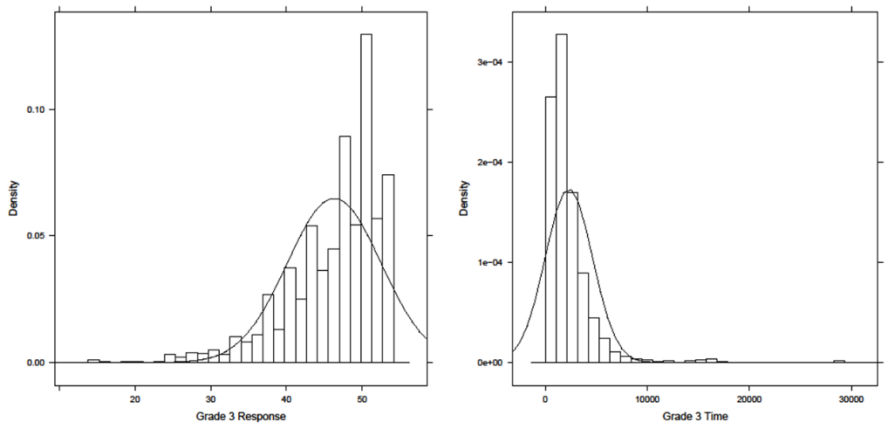

Figure 3. *Histograms of total score (left) and total response time (right) for grade 3.*

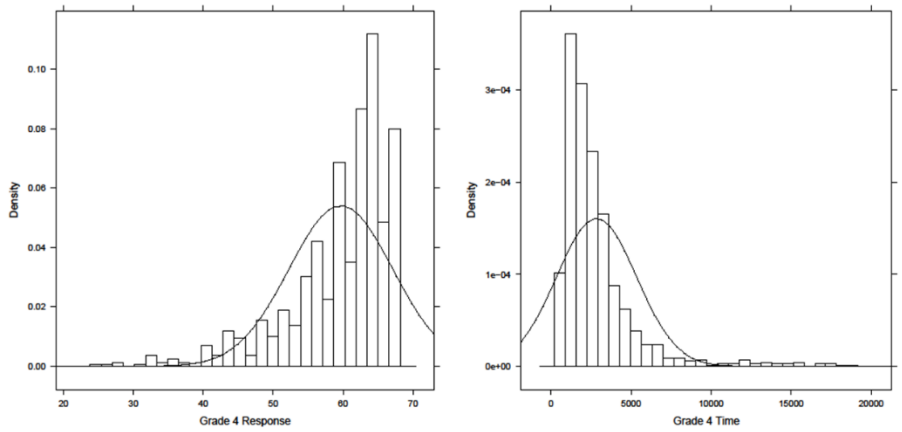

Figure 4. *Histograms of total score (left) and total response time (right) for grade 4.*

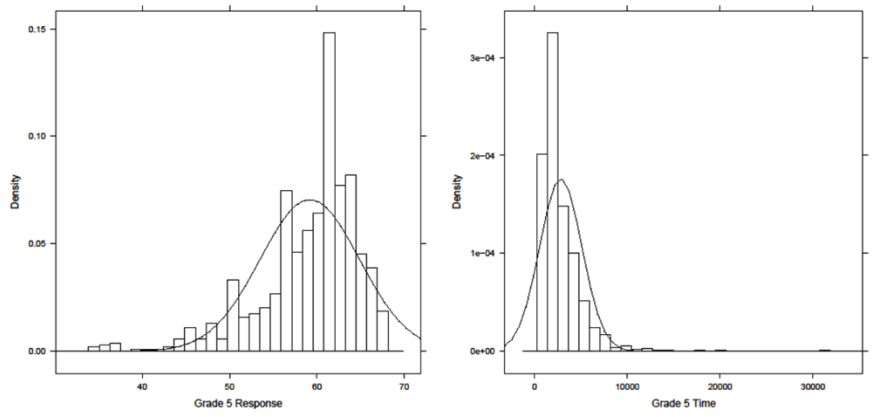

Figure 5. *Histograms of total score (left) and total response time (right) for grade 5.*

The test reliability estimates using Cronbach's coefficient alpha are presented in Table 5. All three grades showed good to excellent internal consistency for response and response time.

Table 5
Coefficient alpha of real data

	Response	**Time**
Grade 3	0.86	0.91
Grade 4	0.89	0.85
Grade 5	0.81	0.86

MCMC Components and Convergence

Table 6 presents the average $\hat{R}$ of parameters $(a, b, \theta, \alpha, \beta, \tau)$ for all grades. The individual value of $\hat{R}$ for parameters (i.e., item parameters will have a $\hat{R}$ for each item and person parameters will have a $\hat{R}$ for each examinee) were categorized into six groups: 1) $1.00 \leq \hat{R} < 1.02$, 2) $1.02 \leq \hat{R} < 1.04$, 3) $1.04 \leq \hat{R} < 1.06$, 4) $1.06 \leq \hat{R} < 1.08$, 5) $1.08 \leq \hat{R} < 1.10$, 6) $1.00 \leq \hat{R}$ (as shown in Table 7). All three grades had average $\hat{R}$ values of 1.00 to 1.07 for parameters with the Markov Chain length of 10,000. While checking the individual $\hat{R}$ value for item and person parameters, the time intensity (β) parameter had a noticeably higher $\hat{R}$ for many items. With the addition of 5,000 iterations, the majority of average $\hat{R}$ values for parameters were close to 1.00 and the individual $\hat{R}$ value for each parameter was categorized into the first two groups. Based on these findings, the data for all three grades were assumed to be converged. The initial length of the Markov Chain and the burn-in period were set at 15,000 and 10,000, respectively. The final length of the Markov Chain for inference making was set at 5,000.

Table 6

Average $\hat{R}$ of each parameter

Length of Markov Chain	Grade	a	b	θ	α	β	τ
10,000	3	1.00	1.00	1.00	1.00	1.05	1.00
	4	1.00	1.00	1.00	1.00	1.07	1.00
	5	1.00	1.00	1.00	1.00	1.02	1.00
15,000	3	1.00	1.00	1.00	1.00	1.00	1.00
	4	1.00	1.00	1.00	1.00	1.00	1.00
	5	1.00	1.00	1.00	1.00	1.01	1.00

Table 7

$\hat{R}$ value of each item and person parameter into six categories

Grade	Category	a	b	θ	α	β	τ
3	$1.00 \leq \hat{R} < 1.02$	54	53	1378	54	54	1378
	$1.02 \leq \hat{R} < 1.04$		1				
	$1.04 \leq \hat{R} < 1.06$						
	$1.06 \leq \hat{R} < 1.08$						
	$1.08 \leq \hat{R} < 1.10$						
	$1.10 \leq \hat{R}$						
4	$1.00 \leq \hat{R} < 1.02$	68	68	1005	68	68	1005
	$1.02 \leq \hat{R} < 1.04$						
	$1.04 \leq \hat{R} < 1.06$						
	$1.06 \leq \hat{R} < 1.08$						
	$1.08 \leq \hat{R} < 1.10$						
	$1.10 \leq \hat{R}$						
5	$1.00 \leq \hat{R} < 1.02$	68	67	888	68	60	888
	$1.02 \leq \hat{R} < 1.04$		1			8	
	$1.04 \leq \hat{R} < 1.06$						
	$1.06 \leq \hat{R} < 1.08$						
	$1.08 \leq \hat{R} < 1.10$						
	$1.10 \leq \hat{R}$						

Note. *Grade 3 contained 54 items and 1,378 examinees. Grade 4 contained 68 items and 1005 examinees. Grade 5 contained 68 items and 888 examinees.*

Parameter Estimates

The item parameter estimates from all three grades are displayed in Tables 8–10. The summaries of item and examinee parameter estimates are presented in Table 11. On average, all three grades had easy items that discriminated well, and examinees took a great amount of time on items. The average examinee abilities for grades 3, 4, and 5 were 0.07, 0.12 and 0.14, respectively, and the average examinee speeds for each grade were -0.10, -0.17, and -0.19, respectively.

Table 8
Item parameter estimates for grade 3

Item	**a**	**b**	**α**	**β**
1	1.05	-2.76	0.91	2.86
2	1.02	-1.02	0.95	2.28
3	2.36	-1.45	1.03	2.16
4	1.08	-2.84	1.09	2.32
5	0.72	-2.57	1.26	1.99
6	2.16	-1.85	1.19	2.02
7	1.65	-2.47	0.60	2.62
8	1.54	-1.43	1.14	2.11
9	1.27	-1.93	1.09	2.13
10	1.75	-1.47	1.14	2.17
11	2.53	-1.69	1.31	1.96
12	3.11	-2.26	1.26	2.01
13	0.65	-2.40	1.03	2.18
14	0.72	-2.84	0.94	3.64
15	1.24	-1.72	0.62	2.93
16	1.62	-1.50	0.92	2.34
17	1.67	-2.17	1.09	2.25
18	1.22	-0.64	1.19	1.96
19	0.73	-0.82	1.16	2.11
20	3.02	-1.86	0.72	2.92
21	1.12	-1.37	0.67	3.36
22	1.87	-1.31	0.62	2.91
23	1.53	-0.28	0.95	3.21
24	1.38	-1.68	1.15	2.55
25	1.78	-1.51	1.15	2.53
26	0.75	0.11	1.11	2.42
27	1.16	-1.26	1.03	2.53
28	3.48	-1.85	1.09	2.53
29	1.17	-2.28	0.99	3.83
30	1.72	-2.68	0.62	2.97
31	1.06	-0.53	0.88	2.50
32	1.98	-2.31	1.01	2.44
33	1.50	-1.50	1.22	2.25
34	1.11	-2.01	1.19	2.30
35	1.18	-3.27	0.77	2.89
36	1.85	-2.18	0.70	3.28
37	1.20	-3.51	0.63	2.65
38	3.07	-2.19	0.88	2.89
39	3.21	-1.99	1.23	2.20
40	2.26	-1.88	1.29	2.10
41	1.31	-3.37	1.15	2.11
42	1.13	-1.67	1.29	1.97
43	1.12	-1.53	1.19	2.00

44	1.89	-2.47	1.01	2.35
45	0.83	-3.64	0.91	3.63
46	0.82	-2.81	0.60	2.82
47	1.22	-1.17	0.84	2.55
48	2.09	-1.48	1.13	2.13
49	0.88	-2.62	1.01	2.30
50	1.68	-1.04	1.13	2.08
51	1.38	-0.94	1.13	2.10
52	2.22	-1.70	1.04	2.23
53	0.84	0.02	0.70	2.90
54	1.12	-1.33	0.66	3.41

Table 9
Item parameter estimates for grade 4

Item	a	b	α	β
1	1.62	-2.25	1.62	3.03
2	1.83	-1.82	1.55	3.16
3	2.44	-2.28	1.11	3.74
4	1.64	-2.30	1.51	3.60
5	2.33	-2.27	0.61	2.91
6	1.90	-2.16	0.82	2.89
7	2.23	-1.56	1.10	2.33
8	1.85	-1.63	1.16	2.34
9	1.07	-1.07	0.98	2.48
10	1.58	-1.34	1.08	2.43
11	0.29	-1.53	1.15	2.31
12	0.66	-2.31	1.11	2.25
13	1.21	-0.62	1.19	2.29
14	1.02	-2.19	1.20	2.04
15	3.17	-1.68	0.91	2.31
16	1.73	-1.25	0.85	3.91
17	1.05	-2.35	0.62	2.75
18	2.47	-1.63	0.87	2.42
19	1.09	-2.00	1.25	2.09
20	2.84	-1.46	1.22	2.06
21	1.78	-1.03	1.29	1.96
22	4.25	-2.14	1.17	2.11
23	1.01	-1.73	0.73	2.69
24	0.87	-1.74	0.67	3.40
25	1.51	-1.46	0.63	2.80
26	1.80	-1.75	0.88	2.95
27	1.27	-1.19	1.10	2.24
28	1.27	-2.03	1.00	2.47
29	1.48	-1.51	1.12	2.29
30	2.07	-1.85	1.07	2.20
31	1.17	-1.88	1.24	1.97
32	2.06	-2.38	1.25	2.18
33	0.77	-1.95	1.19	2.05
34	0.49	-2.86	0.97	2.45
35	1.88	-1.80	0.92	3.70
36	1.14	-1.66	0.65	2.58
37	1.65	-1.38	0.86	2.40
38	2.63	-2.65	0.95	2.29
39	1.80	-1.44	1.14	1.92
40	2.07	-1.10	1.07	2.07
41	2.14	-1.62	0.96	2.38
42	1.38	-1.22	0.89	3.96
43	2.68	-1.72	0.71	2.47

44	1.88	-2.06	1.02	2.21
45	1.56	-1.33	1.21	2.06
46	2.00	-1.69	1.13	2.10
47	1.74	-1.57	1.21	1.96
48	1.59	-1.56	0.90	2.51
49	1.49	-1.43	0.81	3.35
50	2.76	-2.23	0.62	2.61
51	1.08	-1.78	1.13	2.15
52	0.69	-1.29	1.21	1.90
53	3.21	-1.72	0.83	2.71
54	1.77	-2.24	1.17	2.15
55	2.68	-2.58	1.15	1.94
56	0.93	-2.19	1.09	2.04
57	1.45	-2.68	1.10	2.01
58	1.22	-0.86	1.13	2.14
59	1.21	-2.13	0.99	3.48
60	1.65	-1.67	1.05	2.19
61	1.34	-0.69	0.65	2.61
62	1.51	-1.17	0.82	2.64
63	1.41	-1.70	1.14	2.13
64	1.48	-1.34	1.05	2.19
65	1.77	-1.48	1.08	2.01
66	2.35	-0.98	1.18	2.02
67	2.35	-1.72	0.92	2.57
68	1.80	-1.19	0.76	3.26

Table 10
Item parameter estimates for grade 5

Item	a	b	α	β
1	0.92	-1.57	1.50	3.41
2	3.82	-2.30	1.78	2.74
3	0.63	-1.77	1.28	3.86
4	1.89	-1.97	1.37	3.71
5	1.61	-2.02	1.61	3.00
6	1.61	-1.81	0.70	2.57
7	2.18	-1.51	1.01	2.56
8	1.73	-2.02	1.18	2.20
9	2.10	-2.41	1.42	1.88
10	3.00	-1.68	1.12	2.08
11	1.10	-1.74	0.95	2.39
12	1.45	-1.87	0.81	3.46
13	0.81	-1.25	0.64	2.41
14	1.16	-2.28	0.90	2.32
15	0.21	-0.49	1.10	2.16
16	2.60	-1.57	1.14	1.89
17	2.02	-1.59	1.31	2.00
18	1.54	-1.26	1.32	1.95
19	0.52	-0.87	1.20	2.01
20	1.59	-2.01	1.14	2.10
21	0.88	-3.21	1.17	2.05
22	1.88	-1.07	0.94	3.50
23	0.69	-1.71	0.59	2.83
24	2.45	-2.50	0.89	2.55
25	1.90	-2.06	1.24	1.94
26	0.86	-0.78	1.01	1.95
27	2.53	-2.04	0.89	2.31
28	2.27	-2.03	1.32	1.84
29	1.57	-2.39	0.78	2.30
30	0.32	1.35	0.72	3.71
31	1.76	-2.51	0.66	2.55
32	2.43	-1.71	0.86	2.61
33	1.12	-1.13	1.28	2.07
34	1.55	-2.16	1.18	2.05
35	1.19	-0.93	1.29	2.18
36	1.84	-2.39	1.23	2.18
37	0.54	1.00	1.16	2.06
38	2.38	-2.50	1.20	2.17
39	0.65	-1.67	1.01	2.04
40	3.29	-1.65	0.91	3.56
41	1.68	-2.77	0.70	2.39
42	3.32	-1.80	0.93	2.34
43	4.03	-2.06	1.14	1.98

44	1.65	-1.93	1.16	2.20
45	0.27	-1.73	0.96	2.31
46	0.66	-3.25	1.02	3.33
47	1.52	-0.73	0.67	2.66
48	1.90	-2.41	0.86	2.70
49	0.91	-2.47	1.17	2.17
50	1.49	-1.57	1.18	2.12
51	2.69	-2.18	0.89	2.61
52	1.10	0.00	0.86	3.40
53	1.67	-2.22	0.60	2.82
54	2.15	-2.18	0.91	3.14
55	0.81	-2.89	1.22	2.46
56	2.15	-2.46	1.13	2.11
57	2.30	-2.28	1.29	2.21
58	2.64	-2.39	1.25	2.30
59	1.22	-1.86	1.25	2.15
60	2.69	-2.02	1.24	2.34
61	1.40	-1.76	1.08	2.47
62	0.84	-1.52	0.97	3.68
63	1.48	-2.34	0.65	2.79
64	1.56	-1.42	0.79	2.72
65	1.25	-2.13	1.04	2.36
66	1.68	-1.71	1.04	2.45
67	1.63	-0.83	0.91	2.66
68	1.17	-1.63	0.90	3.56

Table 11
Summary of item and examinee parameter estimates

Grade	**a**		**b**		**θ**		**α**		**β**		**τ**	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
3	1.56	0.70	-1.83	0.83	0.07	0.82	0.99	0.21	2.52	0.48	-0.10	0.73
4	1.71	0.69	-1.72	0.48	0.12	0.80	1.03	0.22	2.50	0.53	-0.17	0.55
5	1.65	0.82	-1.77	0.79	0.14	0.74	1.05	0.25	2.52	0.54	-0.19	0.53

Table 12 shows the correlation coefficients among estimated item parameters (*a*, *b*, *α*, and *β*), and Table 13 shows the correlation coefficients between estimated person parameters (θ and τ). The correlations between item parameter estimates ranged from -0.69 to 0.17, and the correlations between person parameters ranged from -0.06 to -0.01. The estimated parameters

had very low positive or low negative correlations, except between the time discrimination (α) and time intensity (β) of grade 3, which had a moderate negative correlation.

Table 12
Correlation coefficients among item parameters

	Grade 3				**Grade 4**				**Grade 5**			
	a	b	α	β	a	b	α	β	a	b	α	β
a												
b	0.02				-0.01				-0.23			
α	0.16	0.14			-0.06	-0.08			0.17	-0.12		
β	-0.17	-0.19	-0.69		-0.01	-0.03	-0.31		-0.13	0.15	-0.24	

Table 13
Correlation coefficients between person parameters

	Grade 3	**Grade 4**	**Grade 5**
	θ	θ	θ
τ	-0.01	-0.05	-0.06

The relationships between the item discrimination (a) and time discrimination (α) parameters, between the item difficulty (b) and time intensity (β) parameters, and between the examinee ability (θ) and examinee speed (τ) parameters are displayed in Figure 6. These parameters (a and α, b and β, θ and τ) were comparable to each other as level-one parameters from response and response time models. For all three comparisons, the parameters had very weak positive or negative relationships between them.

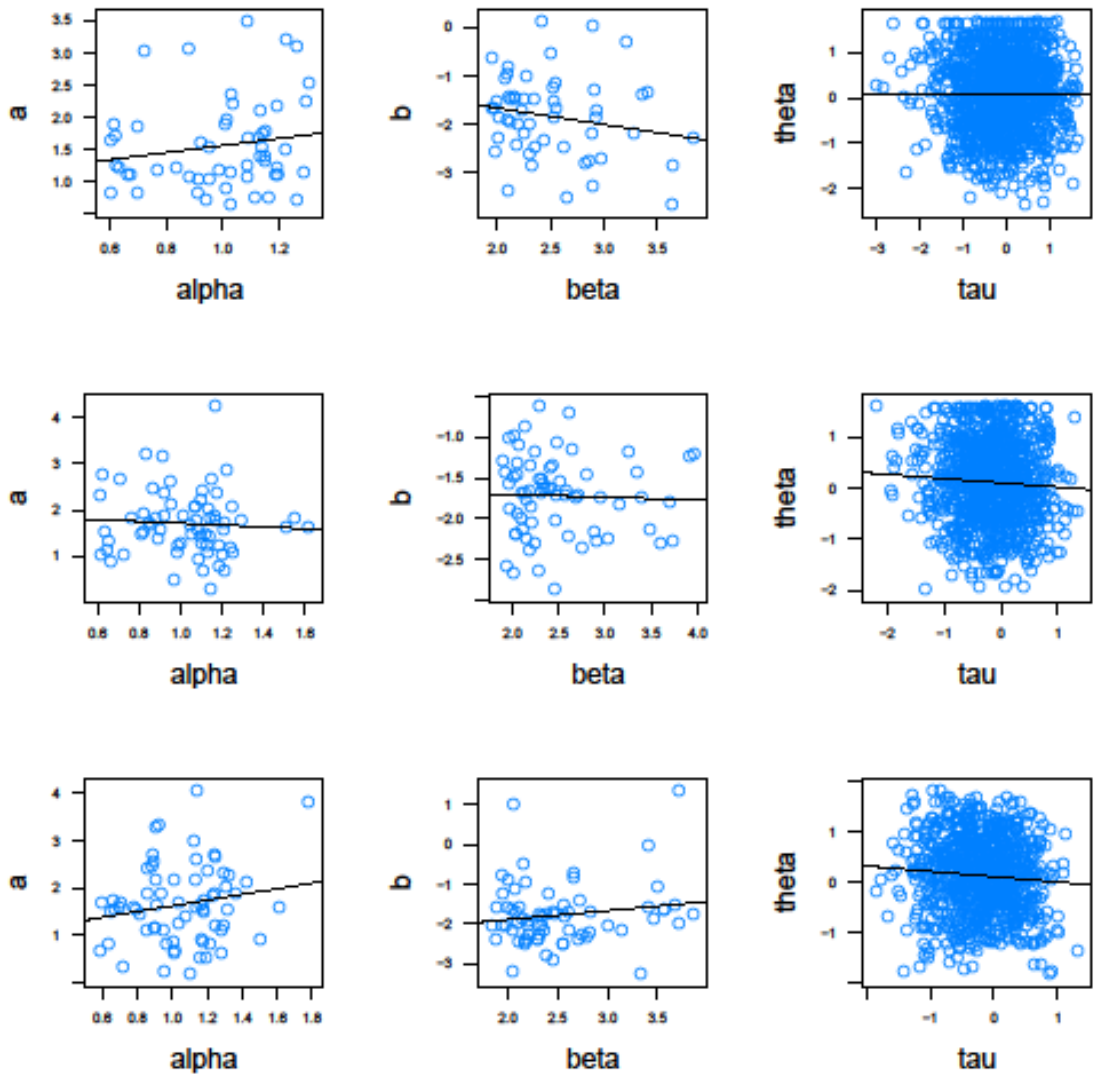

Figure 6. *Scatter plots of estimated parameters between response and response time models for grade 3 (top), grade 4 (center), and grade 5 (bottom).*

The test characteristic curve (TCC) was created by summing each item characteristic curve (ICC) across the ability and speed continua. The TCC for response parameters (a, b, and θ) and response time parameters (α, β, and τ) are shown in Figures 7–9. The vertical axis reflects the expected score on the test for an examinee with a given ability or speed level. Generally, for all grades, the inflection of the curve was at the low level of ability (θ) with response parameters,

and the inflection of the curve was at the high level of speed (τ) with response time parameters. The examinees are expected to have high scores if their ability is near mean of zero. The examinees are expected to have low scores if their speed is near mean of zero. The assessments were easy enough for examinees with below-average abilities to have high expected scores, and examinees who were very fast had high expected scores.

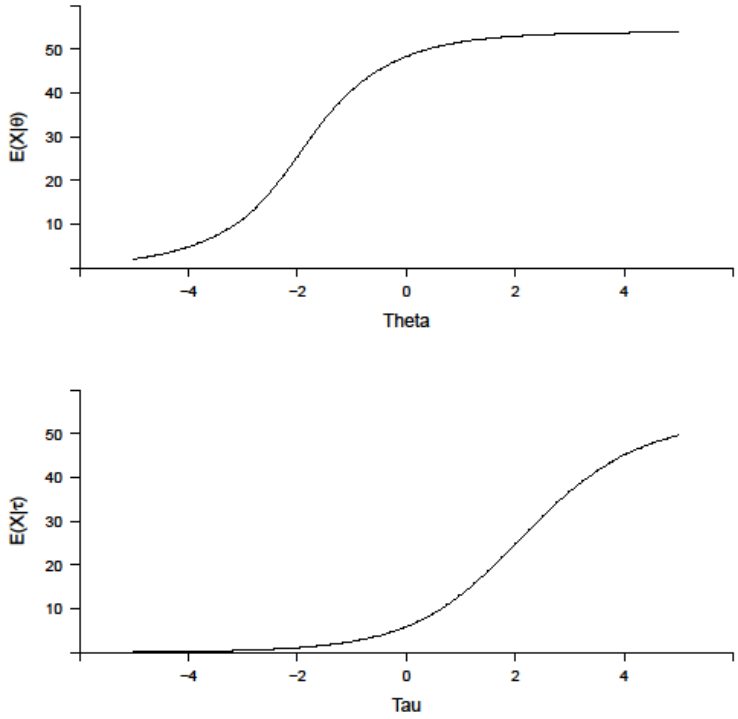

Figure 7. *Test characteristic curve using parameters from responses (top) and response times for grade 3.*

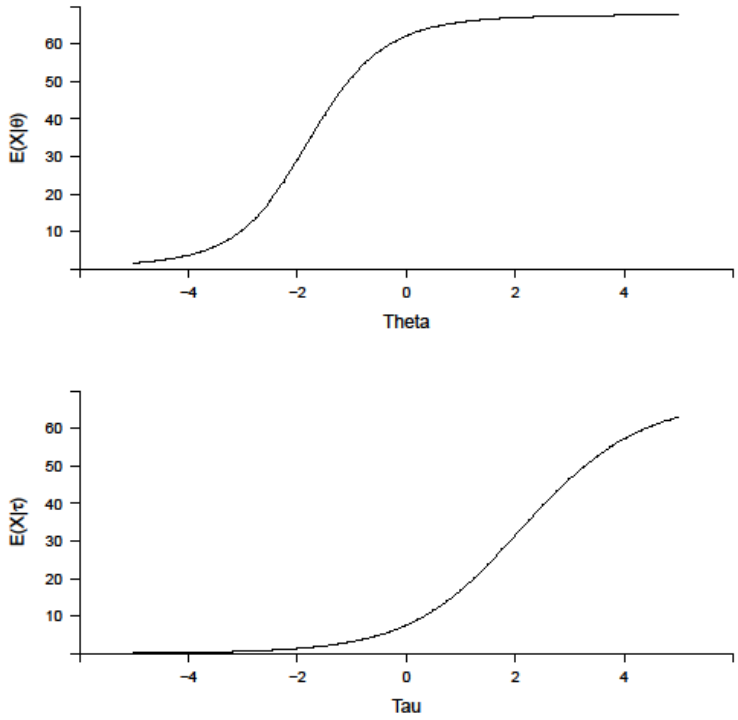

Figure 8. *Test characteristic curve using parameters from responses (top) and response times for grade 4.*

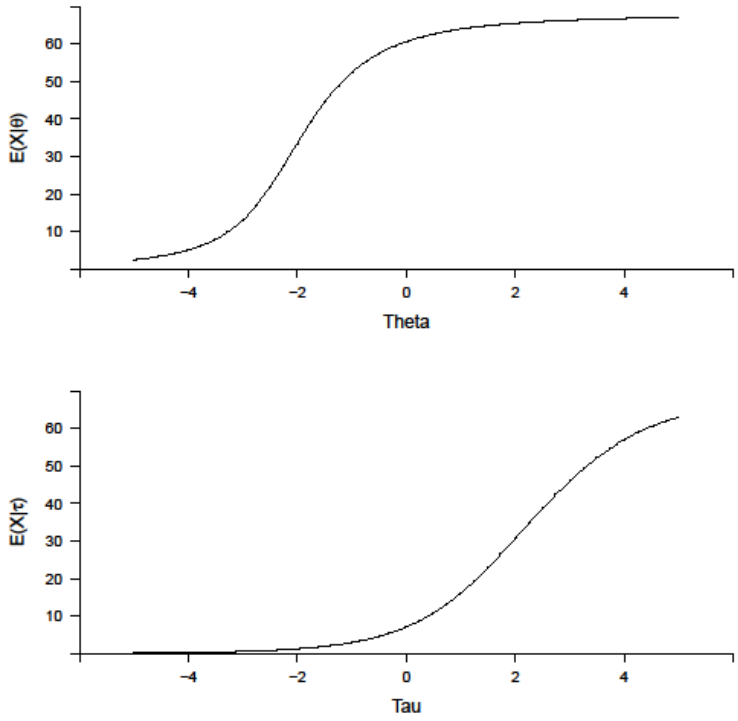

Figure 9. *Test characteristic curve using parameters from responses (top) and response times for grade 5.*

The test information function (TIF) was generated by summing the item information functions (IIF) (as shown in Figure 10–11). Test information is influenced by the quality and the number of test items. The TIF indicates the degree of precision across the ability continuum. For the TIF using response parameters, the test information was maximized around the ability (θ) of -2.0. For the TIF using response time parameters, the amount of information was very low compared with the TIF using response parameters. Information was maximized around a speed (τ) of 2.0.

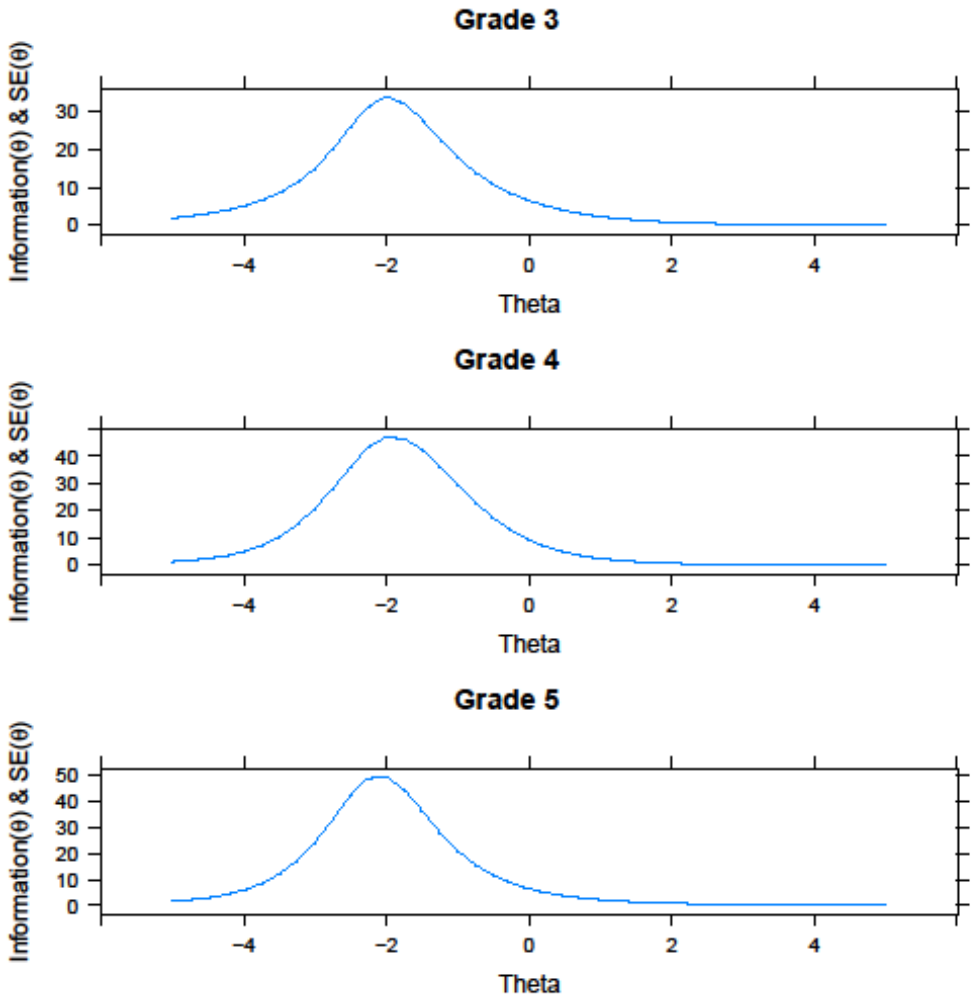

Figure 10. *Test information function (TIF) using response parameters for grade 3 (top), grade 4 (center), and grade 5 (bottom).*

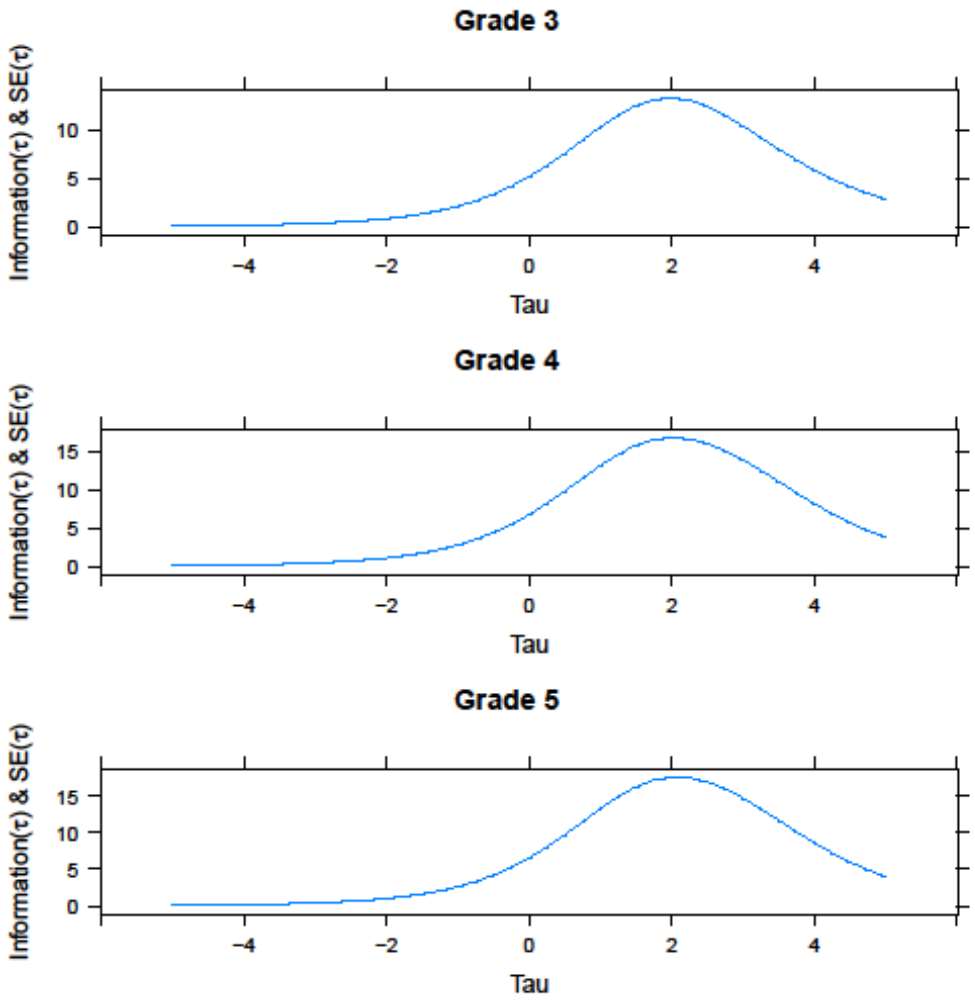

Figure 11. *Test information function (TIF) using response time parameters for grade 3 (top), grade 4 (center), and grade 5 (bottom).*

Study 2

MCMC Components and Convergence

The study applied the MCMC component specifications from Study 1 (i.e., the length of the Markov Chain and the burn-in period). The length of the Markov Chain was set at 15,000 and the burn-in period was set at 10,000. The last 5,000 draws from two separate chains were retained for inference making and the rest of the draws were considered as the burn-in period to be discarded. Tables 14 and 15 present the average $\hat{R}$ of parameters ($a, b, \theta, \alpha, \beta, \tau, \gamma,$ and δ) of all grades for both models. The item parameters (a, b, α, and β) have a $\hat{R}$ value for each item, and person parameters (θ and τ) have a $\hat{R}$ value for each examinee. The $\hat{R}$ values for items or examinees were categorized into six groups: 1) $1.00 \leq \hat{R} < 1.02$, 2) $1.02 \leq \hat{R} < 1.04$, 3) $1.04 \leq \hat{R} < 1.06$, 4) $1.06 \leq \hat{R} < 1.08$, 5) $1.08 \leq \hat{R} < 1.10$, 6) $1.00 \leq \hat{R}$. For all three grades, item and person parameters for both models had an average $\hat{R}$ value near 1.00 in all nine testlet conditions. All parameters had most of their individual $\hat{R}$ values in the first category. The γ-parameter had $\hat{R}$ values near 1.04 or 1.05 for some conditions in grade 5. However, the majority conditions for all three grades had γ- and δ-parameters' $\hat{R}$ values near 1.00. Based on these results, convergence was assumed for all conditions.

Table 14

Average $\hat{R}$ value of each parameter to check convergence for HTRT model

Grade	Condition	a	b	θ	γ	α	β	τ	δ
3	0.25 & 0.25	1.00	1.00	1.00	1.02	1.00	1.01	1.00	1.00
	0.25 & 0.50	1.00	1.00	1.00	1.02	1.00	1.01	1.00	1.00
	0.25 & 1.00	1.00	1.00	1.00	1.02	1.00	1.01	1.00	1.00
	0.50 & 0.25	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.00
	0.50 & 0.50	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.00
	0.50 & 1.00	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.00
	1.00 & 0.25	1.00	1.01	1.00	1.01	1.00	1.01	1.00	1.00
	1.00 & 0.50	1.00	1.01	1.00	1.01	1.00	1.01	1.00	1.00
	1.00 & 1.00	1.00	1.01	1.00	1.01	1.00	1.01	1.00	1.00
4	0.25 & 0.25	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.00
	0.25 & 0.50	1.00	1.00	1.00	1.02	1.00	1.01	1.00	1.00
	0.25 & 1.00	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.00
	0.50 & 0.25	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.00
	0.50 & 0.50	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.00
	0.50 & 1.00	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.00
	1.00 & 0.25	1.00	1.01	1.00	1.01	1.00	1.01	1.00	1.00
	1.00 & 0.50	1.00	1.01	1.00	1.01	1.00	1.01	1.00	1.00
	1.00 & 1.00	1.00	1.01	1.00	1.01	1.00	1.01	1.00	1.00
5	0.25 & 0.25	1.00	1.00	1.00	1.04	1.00	1.01	1.00	1.00
	0.25 & 0.50	1.00	1.00	1.00	1.05	1.00	1.01	1.00	1.00
	0.25 & 1.00	1.00	1.00	1.00	1.04	1.00	1.01	1.00	1.00
	0.50 & 0.25	1.00	1.00	1.00	1.01	1.00	1.02	1.00	1.00
	0.50 & 0.50	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.00
	0.50 & 1.00	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.00
	1.00 & 0.25	1.00	1.01	1.00	1.01	1.00	1.01	1.00	1.00
	1.00 & 0.50	1.00	1.01	1.00	1.01	1.00	1.01	1.00	1.00
	1.00 & 1.00	1.00	1.00	1.00	1.01	1.00	1.02	1.00	1.00

Table 15

Average $\hat{R}$ value of each parameter to check convergence for Hierarchical Framework model

Grade	Condition	a	b	θ	α	β	τ
3	0.25 & 0.25	1.00	1.00	1.00	1.00	1.01	1.00
	0.25 & 0.50	1.00	1.00	1.00	1.00	1.01	1.00
	0.25 & 1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.50 & 0.25	1.00	1.00	1.00	1.00	1.01	1.00
	0.50 & 0.50	1.00	1.00	1.00	1.00	1.01	1.00
	0.50 & 1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1.00 & 0.25	1.00	1.00	1.00	1.00	1.01	1.00
	1.00 & 0.50	1.00	1.00	1.00	1.00	1.01	1.00
	1.00 & 1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.25 & 0.25	1.00	1.00	1.00	1.00	1.01	1.00
	0.25 & 0.50	1.00	1.00	1.00	1.00	1.01	1.00
	0.25 & 1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.50 & 0.25	1.00	1.00	1.00	1.00	1.02	1.00
	0.50 & 0.50	1.00	1.00	1.00	1.00	1.01	1.00
	0.50 & 1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1.00 & 0.25	1.00	1.00	1.00	1.00	1.01	1.00
	1.00 & 0.50	1.00	1.00	1.00	1.00	1.01	1.00
	1.00 & 1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	0.25 & 0.25	1.00	1.00	1.00	1.00	1.01	1.00
	0.25 & 0.50	1.00	1.00	1.00	1.00	1.01	1.00
	0.25 & 1.00	1.00	1.00	1.00	1.00	1.01	1.00
	0.50 & 0.25	1.00	1.00	1.00	1.00	1.01	1.00
	0.50 & 0.50	1.00	1.00	1.00	1.00	1.01	1.00
	0.50 & 1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1.00 & 0.25	1.00	1.00	1.00	1.00	1.01	1.00
	1.00 & 0.50	1.00	1.00	1.00	1.00	1.00	1.00
	1.00 & 1.00	1.00	1.00	1.00	1.00	1.01	1.00

DIC Comparison

Table 16 compares DIC values of the HTRT model and the Hierarchical Framework model on various testlet conditions. Overall, the HTRT model showed better fit by having lower DIC values than the Hierarchical Framework model in all conditions. The HTRT model had very

stable DIC values regardless of the presence of testlet variance. However, DIC values for the Hierarchical Framework model increased as the variances of γ and/or δ increased.

Table 16
DIC values from both response time models for nine testlet conditions

Grade	Condition	HTRT	Hierarchical Framework
3	0.25 & 0.25	648035.89	664742.92
	0.25 & 0.50	649779.87	677890.61
	0.25 & 1.00	651919.49	697958.74
	0.50 & 0.25	647749.66	666538.35
	0.50 & 0.50	649570.71	679681.63
	0.50 & 1.00	651654.44	699739.06
	1.00 & 0.25	648253.24	671184.64
	1.00 & 0.50	650075.14	684456.03
	1.00 & 1.00	652167.45	704478.84
4	0.25 & 0.25	594949.25	610365.55
	0.25 & 0.50	596172.54	622192.16
	0.25 & 1.00	598532.30	641682.27
	0.50 & 0.25	594708.14	612130.73
	0.50 & 0.50	595915.86	623969.86
	0.50 & 1.00	598301.08	643474.41
	1.00 & 0.25	595588.34	616850.77
	1.00 & 0.50	596782.93	628690.04
	1.00 & 1.00	599159.61	648156.41
5	0.25 & 0.25	527802.14	541219.42
	0.25 & 0.50	529216.05	552210.17
	0.25 & 1.00	531379.38	569971.62
	0.50 & 0.25	527833.35	542698.58
	0.50 & 0.50	529336.54	553705.04
	0.50 & 1.00	531460.28	571447.92
	1.00 & 0.25	528943.98	546484.51
	1.00 & 0.50	530342.30	557391.57
	1.00 & 1.00	532588.17	575215.74

Relationship between response parameters and response time parameters

The relationship between the item discrimination (a) and time discrimination (α) parameters, between the item difficulty (b) and time intensity (β) parameters, and between the examinee ability (θ) and examinee speed (τ) parameters for both response time models of all

three grades are displayed in Figures A1– A9 in the Appendix. These parameters (a and α, b and β, θ and τ) were comparable with each other as level-one parameters from the response (a, b, and θ) and response time (α, β, and τ) models.

Overall, both the HTRT model and the Hierarchical Framework model showed very weak positive or negative relationships between response and response time parameters throughout all grades. These outcomes were similar to the relationships found from real data.

Item Parameter Recovery

The summaries of item and examinee parameter estimates are presented in Tables 17 and 18. Generally, the parameter estimates from the HTRT model were similar across different testlet conditions. The presence of testlet variances did not affect parameter estimation when the HTRT model was used. With the Hierarchical Framework model, however, the amount of shared variance on testlet parameters (γ and δ) had a clear impact on the parameter estimation. The average parameter estimates showed either an increase or decrease with the presence of testlet variances for item parameters.

Table 17

Means and standard deviations of parameter estimates for HTRT model

		HTRT							
Grade	Condition	a	b	θ	γ	α	β	τ	δ
3	0.25 & 0.25	1.52 (0.68)	-1.88 (0.77)	0.07 (0.73)	0.25 (0.01)	0.99 (0.21)	2.50 (0.48)	-0.10 (0.71)	0.25 (0.01)
	0.25 & 0.50	1.52 (0.66)	-1.88 (0.77)	0.07 (0.73)	0.25 (0.02)	0.99 (0.21)	2.51 (0.49)	-0.10 (0.71)	0.49 (0.02)
	0.25 & 1.00	1.52 (0.66)	-1.87 (0.77)	0.07 (0.73)	0.25 (0.02)	0.99 (0.21)	2.52 (0.48)	-0.10 (0.69)	0.97 (0.04)
	0.50 & 0.25	1.53 (0.67)	-1.87 (0.76)	0.07 (0.72)	0.48 (0.09)	0.99 (0.21)	2.50 (0.48)	-0.10 (0.71)	0.25 (0.01)
	0.50 & 0.50	1.53 (0.65)	-1.87 (0.76)	0.07 (0.72)	0.49 (0.09)	0.99 (0.21)	2.51 (0.49)	-0.10 (0.71)	0.49 (0.02)
	0.50 & 1.00	1.53 (0.64)	-1.87 (0.76)	0.07 (0.72)	0.49 (0.09)	0.99 (0.21)	2.52 (0.48)	-0.10 (0.69)	0.97 (0.04)
	1.00 & 0.25	1.52 (0.66)	-1.86 (0.80)	0.07 (0.70)	1.01 (0.23)	0.99 (0.21)	2.50 (0.48)	-0.10 (0.71)	0.25 (0.01)
	1.00 & 0.50	1.52 (0.66)	-1.86 (0.79)	0.07 (0.70)	1.01 (0.22)	0.99 (0.21)	2.51 (0.49)	-0.10 (0.71)	0.49 (0.02)
	1.00 & 1.00	1.52 (0.66)	-1.86 (0.80)	0.07 (0.70)	1.01 (0.22)	0.99 (0.21)	2.52 (0.48)	-0.10 (0.69)	0.97 (0.04)
4	0.25 & 0.25	1.67 (0.67)	-1.79 (0.45)	0.12 (0.72)	0.24 (0.04)	1.03 (0.22)	2.48 (0.53)	-0.17 (0.55)	0.25 (0.01)
	0.25 & 0.50	1.67 (0.67)	-1.78 (0.46)	0.12 (0.72)	0.23 (0.03)	1.03 (0.22)	2.48 (0.53)	-0.17 (0.55)	0.50 (0.02)
	0.25 & 1.00	1.67 (0.67)	-1.79 (0.46)	0.12 (0.72)	0.23 (0.03)	1.03 (0.22)	2.49 (0.53)	-0.17 (0.55)	1.00 (0.05)
	0.50 & 0.25	1.68 (0.68)	-1.79 (0.45)	0.12 (0.71)	0.45 (0.07)	1.03 (0.22)	2.48 (0.53)	-0.17 (0.55)	0.25 (0.01)
	0.50 & 0.50	1.68 (0.66)	-1.79 (0.46)	0.12 (0.71)	0.45 (0.06)	1.03 (0.22)	2.48 (0.53)	-0.17 (0.55)	0.50 (0.02)
	0.50 & 1.00	1.68 (0.66)	-1.79 (0.46)	0.12 (0.71)	0.45 (0.06)	1.03 (0.22)	2.49 (0.53)	-0.17 (0.55)	1.00 (0.05)
	1.00 & 0.25	1.67 (0.65)	-1.76 (0.46)	0.12 (0.70)	0.87 (0.09)	1.03 (0.22)	2.48 (0.53)	-0.17 (0.55)	0.25 (0.01)
	1.00 & 0.50	1.68 (0.67)	-1.79 (0.46)	0.12 (0.70)	0.87 (0.08)	1.03 (0.22)	2.48 (0.53)	-0.17 (0.55)	0.50 (0.02)
	1.00 & 1.00	1.67 (0.66)	-1.79 (0.47)	0.12 (0.70)	0.87 (0.08)	1.03 (0.22)	2.49 (0.53)	-0.17 (0.55)	1.00 (0.05)
5	0.25 & 0.25	1.68 (0.82)	-1.81 (0.75)	0.14 (0.64)	0.22 (0.02)	1.06 (0.25)	2.51 (0.54)	-0.19 (0.53)	0.25 (0.01)
	0.25 & 0.50	1.67 (0.81)	-1.81 (0.75)	0.14 (0.64)	0.23 (0.03)	1.05 (0.25)	2.51 (0.54)	-0.19 (0.52)	0.50 (0.01)
	0.25 & 1.00	1.69 (0.82)	-1.80 (0.75)	0.14 (0.64)	0.22 (0.04)	1.05 (0.25)	2.52 (0.53)	-0.19 (0.52)	1.00 (0.04)
	0.50 & 0.25	1.69 (0.83)	-1.82 (0.75)	0.14 (0.64)	0.44 (0.07)	1.06 (0.25)	2.50 (0.54)	-0.19 (0.53)	0.25 (0.01)
	0.50 & 0.50	1.67 (0.81)	-1.82 (0.75)	0.14 (0.64)	0.44 (0.08)	1.05 (0.25)	2.51 (0.54)	-0.19 (0.52)	0.50 (0.01)
	0.50 & 1.00	1.68 (0.81)	-1.82 (0.75)	0.14 (0.64)	0.44 (0.08)	1.05 (0.25)	2.52 (0.53)	-0.19 (0.52)	1.00 (0.04)
	1.00 & 0.25	1.66 (0.77)	-1.77 (0.77)	0.14 (0.65)	0.81 (0.14)	1.06 (0.25)	2.51 (0.54)	-0.19 (0.53)	0.25 (0.01)
	1.00 & 0.50	1.66 (0.76)	-1.77 (0.76)	0.14 (0.65)	0.82 (0.15)	1.05 (0.25)	2.51 (0.54)	-0.19 (0.52)	0.50 (0.01)
	1.00 & 1.00	1.66 (0.76)	-1.77 (0.76)	0.14 (0.65)	0.81 (0.15)	1.05 (0.25)	2.52 (0.53)	-0.19 (0.52)	1.00 (0.04)

Table 18

Means and standard deviations of parameter estimates for Hierarchical Framework model

		Hierarchical Framework					
Grade	Condition	a	b	θ	α	β	τ
3	0.25 & 0.25	1.44 (0.56)	-1.82 (0.74)	0.07 (0.75)	0.90 (0.17)	2.51 (0.48)	-0.10 (0.74)
	0.25 & 0.50	1.44 (0.55)	-1.82 (0.73)	0.07 (0.75)	0.83 (0.14)	2.52 (0.48)	-0.10 (0.76)
	0.25 & 1.00	1.44 (0.55)	-1.82 (0.73)	0.07 (0.75)	0.73 (0.11)	2.53 (0.48)	-0.10 (0.81)
	0.50 & 0.25	1.39 (0.51)	-1.78 (0.71)	0.07 (0.76)	0.90 (0.17)	2.51 (0.48)	-0.10 (0.74)
	0.50 & 0.50	1.38 (0.49)	-1.78 (0.71)	0.07 (0.76)	0.83 (0.14)	2.52 (0.48)	-0.10 (0.76)
	0.50 & 1.00	1.39 (0.50)	-1.78 (0.71)	0.07 (0.76)	0.73 (0.11)	2.53 (0.48)	-0.10 (0.81)
	1.00 & 0.25	1.29 (0.46)	-1.71 (0.72)	0.07 (0.78)	0.90 (0.17)	2.51 (0.48)	-0.10 (0.74)
	1.00 & 0.50	1.29 (0.45)	-1.72 (0.72)	0.07 (0.78)	0.83 (0.14)	2.52 (0.48)	-0.10 (0.76)
	1.00 & 1.00	1.29 (0.46)	-1.71 (0.72)	0.07 (0.78)	0.73 (0.11)	2.53 (0.48)	-0.10 (0.81)
4	0.25 & 0.25	1.59 (0.61)	-1.74 (0.43)	0.12 (0.73)	0.93 (0.19)	2.49 (0.53)	-0.17 (0.58)
	0.25 & 0.50	1.59 (0.61)	-1.73 (0.44)	0.12 (0.73)	0.86 (0.18)	2.49 (0.52)	-0.17 (0.59)
	0.25 & 1.00	1.59 (0.61)	-1.74 (0.44)	0.12 (0.73)	0.76 (0.16)	2.51 (0.53)	-0.17 (0.65)
	0.50 & 0.25	1.52 (0.53)	-1.71 (0.41)	0.12 (0.74)	0.93 (0.19)	2.49 (0.53)	-0.17 (0.58)
	0.50 & 0.50	1.52 (0.51)	-1.71 (0.43)	0.12 (0.74)	0.86 (0.18)	2.49 (0.52)	-0.17 (0.59)
	0.50 & 1.00	1.52 (0.52)	-1.71 (0.42)	0.11 (0.74)	0.76 (0.16)	2.51 (0.53)	-0.17 (0.65)
	1.00 & 0.25	1.40 (0.45)	-1.64 (0.41)	0.11 (0.76)	0.93 (0.19)	2.49 (0.53)	-0.17 (0.58)
	1.00 & 0.50	1.40 (0.44)	-1.63 (0.41)	0.11 (0.76)	0.86 (0.18)	2.49 (0.52)	-0.17 (0.59)
	1.00 & 1.00	1.40 (0.44)	-1.64 (0.41)	0.11 (0.76)	0.76 (0.16)	2.51 (0.53)	-0.17 (0.65)
5	0.25 & 0.25	1.61 (0.76)	-1.75 (0.72)	0.13 (0.65)	0.95 (0.22)	2.51 (0.54)	-0.19 (0.55)
	0.25 & 0.50	1.60 (0.74)	-1.75 (0.72)	0.14 (0.65)	0.88 (0.21)	2.52 (0.54)	-0.19 (0.56)
	0.25 & 1.00	1.61 (0.75)	-1.75 (0.72)	0.13 (0.65)	0.77 (0.20)	2.54 (0.53)	-0.19 (0.60)
	0.50 & 0.25	1.53 (0.70)	-1.73 (0.70)	0.13 (0.66)	0.95 (0.22)	2.51 (0.54)	-0.19 (0.55)
	0.50 & 0.50	1.52 (0.68)	-1.74 (0.71)	0.13 (0.66)	0.88 (0.21)	2.52 (0.54)	-0.19 (0.56)
	0.50 & 1.00	1.53 (0.69)	-1.74 (0.70)	0.13 (0.66)	0.77 (0.20)	2.54 (0.53)	-0.19 (0.60)
	1.00 & 0.25	1.40 (0.60)	-1.66 (0.70)	0.13 (0.70)	0.95 (0.22)	2.51 (0.54)	-0.19 (0.55)
	1.00 & 0.50	1.41 (0.60)	-1.66 (0.70)	0.13 (0.70)	0.88 (0.21)	2.52 (0.54)	-0.19 (0.56)
	1.00 & 1.00	1.40 (0.60)	-1.66 (0.69)	0.13 (0.70)	0.77 (0.20)	2.54 (0.53)	-0.19 (0.60)

For each replication, the parameter estimates were compared with their true values. Tables 19 and 20 and Figures 12–17 contain the results for marginal bias (systematic error) of item and person parameters for both models. These biases are calculated by averaging parameter bias within a replication, and calculating the mean and standard deviation of biases over the replications. The time intensity (β), the examinee ability (θ), and the examinee speed (τ) parameters did not show much difference between the two models. In general, the HTRT model showed smaller biases with the item discrimination (a), the item difficulty (b), and the time discrimination (α) parameters. For the Hierarchical Framework model, the magnitude of biases for the item discrimination (a), the item difficulty (b), and the time discrimination (α) parameters increased as the amount of variance for testlet parameters (γ and δ) increased. For the HTRT model, parameter estimation was not affected by the amount of variance for testlet parameters (γ and δ). For the Hierarchical Framework model, the item difficulty (b) parameter was somewhat biased but mostly positive, and the item discrimination (a) and the time discrimination (α) parameters were negatively biased. The amount of biases for the item difficulty (b) parameter and the item discrimination (a) were affected by the amount of variance in the γ-parameter. The amount of bias for the time discrimination (α) parameter was influenced by the amount of variance in the δ-parameter.

Table 19

Mean bias for response parameter estimates

		HTRT			Hierarchical Framework		
Grade	Condition	*a*	*b*	θ	*a*	*b*	θ
3	0.25 & 0.25	-0.03	-0.05	0.00	-0.12	0.01	0.00
	0.25 & 0.50	-0.04	-0.04	0.00	-0.12	0.01	0.00
	0.25 & 1.00	-0.03	-0.04	0.00	-0.12	0.01	0.00
	0.50 & 0.25	-0.02	-0.04	0.00	-0.17	0.05	0.00
	0.50 & 0.50	-0.03	-0.04	0.00	-0.17	0.05	0.00
	0.50 & 1.00	-0.03	-0.04	0.00	-0.17	0.05	0.00
	1.00 & 0.25	-0.04	-0.03	0.00	-0.26	0.12	-0.01
	1.00 & 0.50	-0.04	-0.03	0.00	-0.27	0.12	-0.01
	1.00 & 1.00	-0.04	-0.03	0.00	-0.27	0.12	-0.01
4	0.25 & 0.25	-0.04	-0.07	0.00	-0.12	-0.02	0.00
	0.25 & 0.50	-0.04	-0.06	0.00	-0.11	-0.01	0.00
	0.25 & 1.00	-0.04	-0.07	0.00	-0.12	-0.02	0.00
	0.50 & 0.25	-0.02	-0.07	0.00	-0.19	0.01	0.00
	0.50 & 0.50	-0.03	-0.07	0.00	-0.19	0.01	-0.01
	0.50 & 1.00	-0.02	-0.07	0.00	-0.19	0.01	-0.01
	1.00 & 0.25	-0.04	-0.04	0.00	-0.31	0.08	-0.01
	1.00 & 0.50	-0.03	-0.03	0.00	-0.31	0.09	-0.01
	1.00 & 1.00	-0.04	-0.04	0.00	-0.31	0.08	-0.01
5	0.25 & 0.25	0.03	-0.03	0.00	-0.04	0.03	0.00
	0.25 & 0.50	0.02	-0.04	0.00	-0.06	0.02	0.00
	0.25 & 1.00	0.03	-0.03	0.00	-0.04	0.03	0.00
	0.50 & 0.25	0.04	-0.04	0.00	-0.12	0.04	0.00
	0.50 & 0.50	0.02	-0.05	0.00	-0.14	0.04	0.00
	0.50 & 1.00	0.03	-0.05	0.00	-0.13	0.04	0.00
	1.00 & 0.25	0.01	0.00	0.00	-0.25	0.12	-0.01
	1.00 & 0.50	0.01	0.00	0.00	-0.24	0.12	-0.01
	1.00 & 1.00	0.00	0.00	0.00	-0.25	0.11	-0.01

Table 20

Mean bias for response time parameter estimates

		HTRT			Hierarchical Framework		
Grade	Condition	α	β	τ	α	β	τ
3	0.25 & 0.25	0.00	-0.01	0.01	-0.10	-0.01	0.01
	0.25 & 0.50	0.00	-0.01	0.01	-0.17	0.00	0.01
	0.25 & 1.00	0.00	0.00	0.01	-0.26	0.02	0.01
	0.50 & 0.25	0.00	-0.01	0.01	-0.10	-0.01	0.01
	0.50 & 0.50	0.00	-0.01	0.01	-0.17	0.00	0.01
	0.50 & 1.00	0.00	0.00	0.01	-0.26	0.01	0.01
	1.00 & 0.25	0.00	-0.01	0.01	-0.10	-0.01	0.01
	1.00 & 0.50	0.00	-0.01	0.01	-0.17	0.00	0.01
	1.00 & 1.00	0.00	0.00	0.01	-0.26	0.01	0.00
4	0.25 & 0.25	0.00	-0.02	0.00	-0.10	-0.01	0.00
	0.25 & 0.50	0.00	-0.01	0.00	-0.17	0.00	0.00
	0.25 & 1.00	0.00	-0.01	0.00	-0.27	0.01	0.00
	0.50 & 0.25	0.00	-0.02	0.00	-0.10	-0.01	0.00
	0.50 & 0.50	0.00	-0.01	0.00	-0.17	0.00	0.00
	0.50 & 1.00	0.00	-0.01	0.00	-0.27	0.01	0.00
	1.00 & 0.25	0.00	-0.02	0.00	-0.10	-0.01	0.00
	1.00 & 0.50	0.00	-0.02	0.00	-0.17	0.00	0.00
	1.00 & 1.00	0.00	-0.01	0.00	-0.27	0.01	0.00
5	0.25 & 0.25	0.00	-0.02	0.00	-0.10	-0.01	0.00
	0.25 & 0.50	0.00	-0.01	0.00	-0.17	0.00	0.00
	0.25 & 1.00	0.00	-0.01	0.00	-0.28	0.01	0.00
	0.50 & 0.25	0.00	-0.02	0.00	-0.10	-0.01	0.00
	0.50 & 0.50	0.00	-0.01	0.00	-0.17	0.00	0.00
	0.50 & 1.00	0.00	-0.01	0.00	-0.28	0.01	0.00
	1.00 & 0.25	0.00	-0.02	0.00	-0.10	-0.01	0.00
	1.00 & 0.50	0.00	-0.01	0.00	-0.17	0.00	0.00
	1.00 & 1.00	0.00	-0.01	0.00	-0.28	0.01	0.00

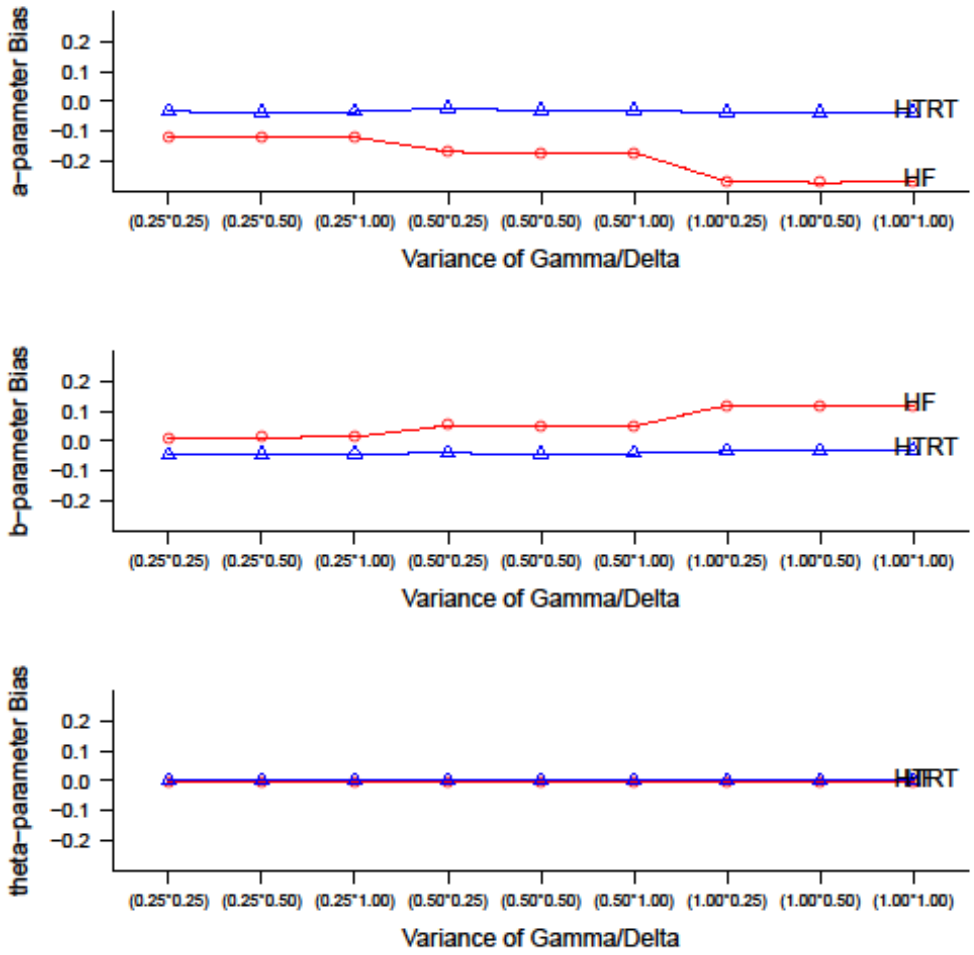

Figure 12. *Marginal bias in the recovery of response parameters for grade 3. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

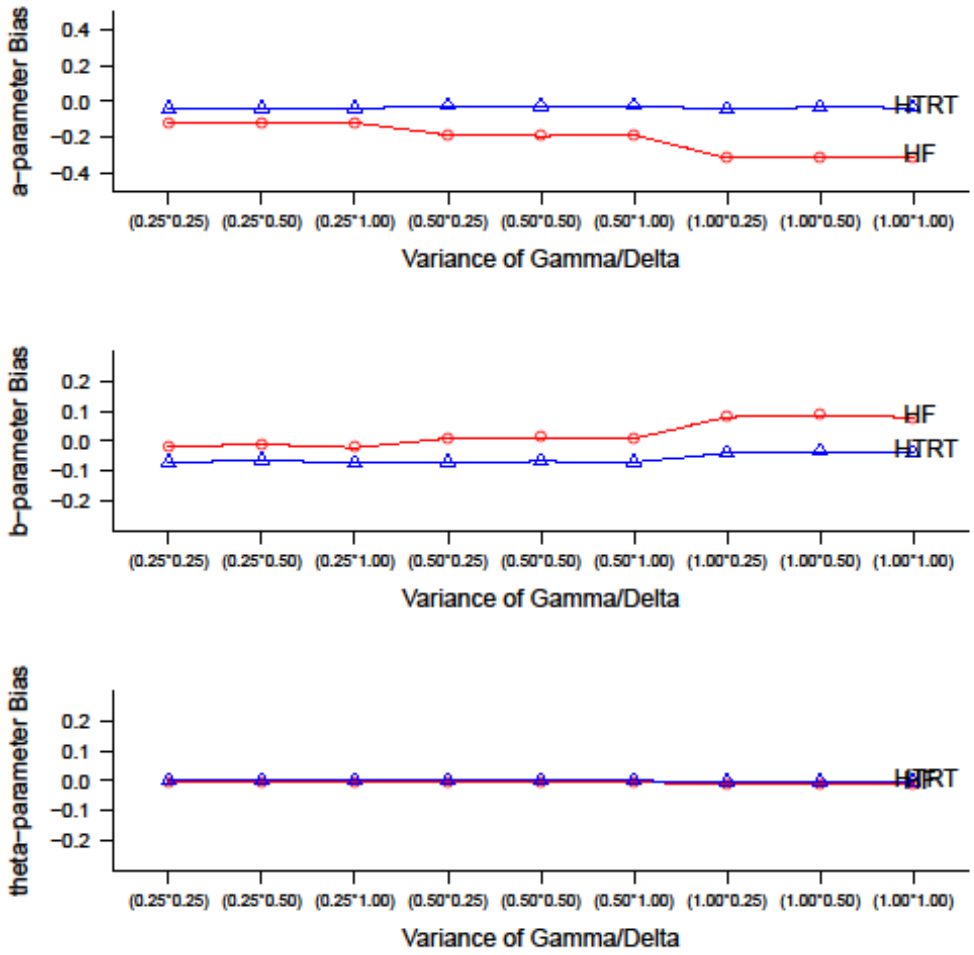

Figure 13. *Marginal bias in the recovery of response parameters for grade 4. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

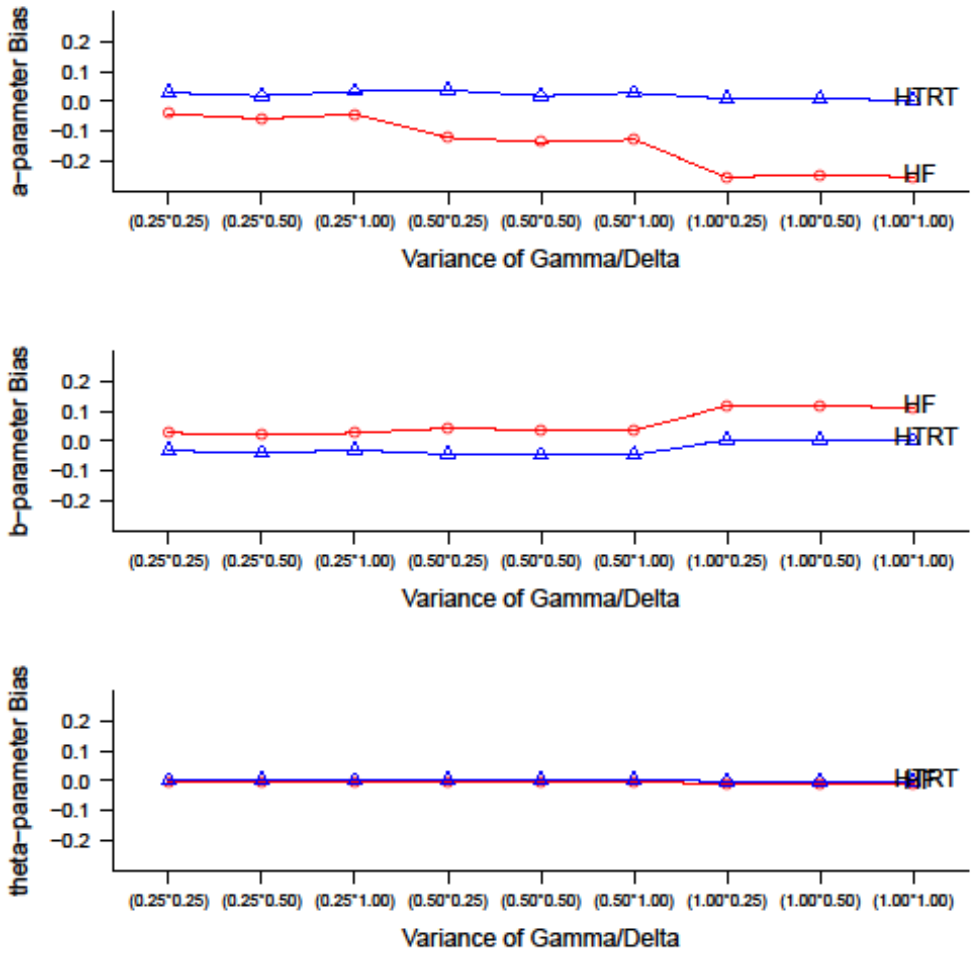

Figure 14. *Marginal bias in the recovery of response parameters for grade 5. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

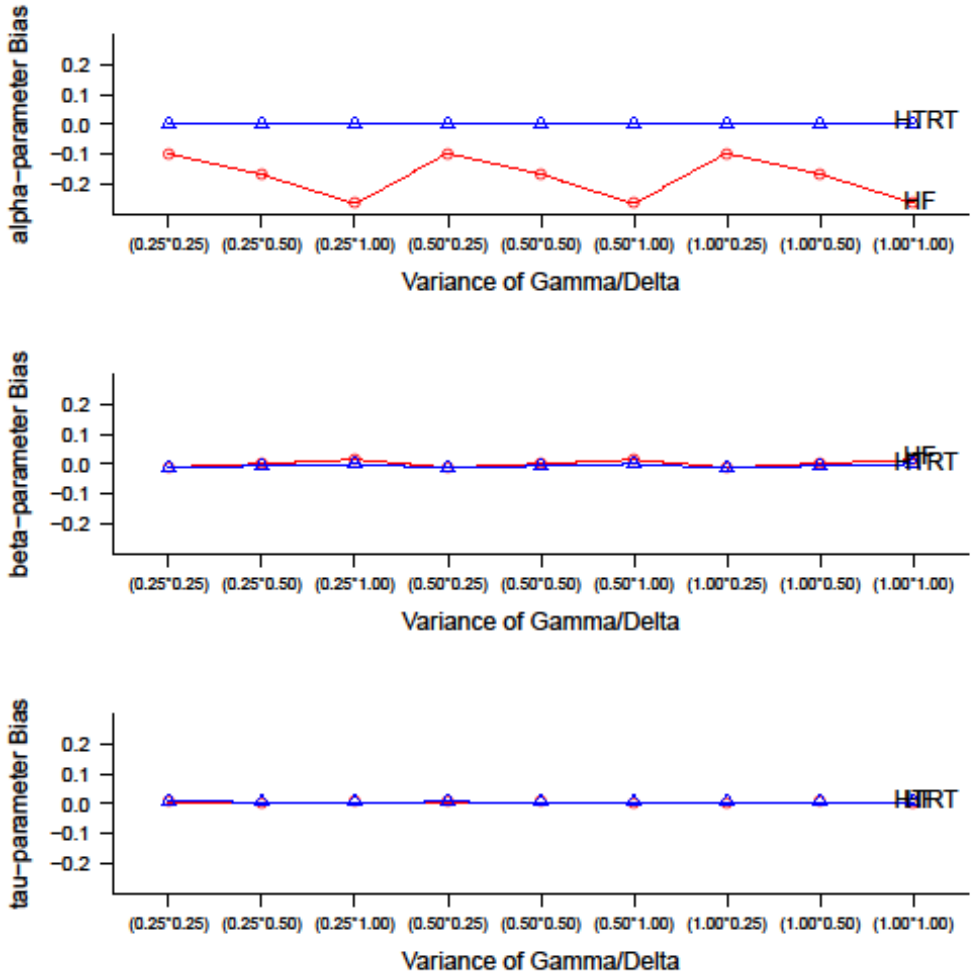

Figure 15. *Marginal bias in the recovery of response time parameters for grade 3. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

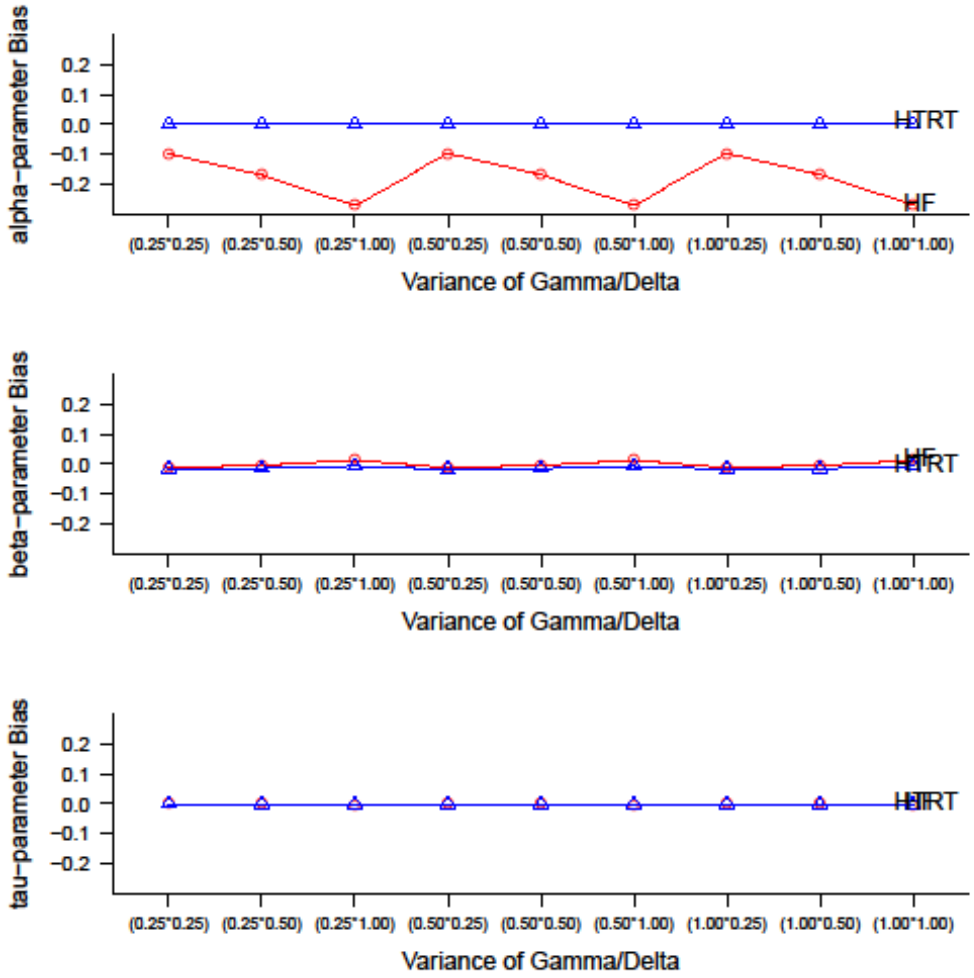

Figure 16. *Marginal bias in the recovery of response time parameters for grade 4. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

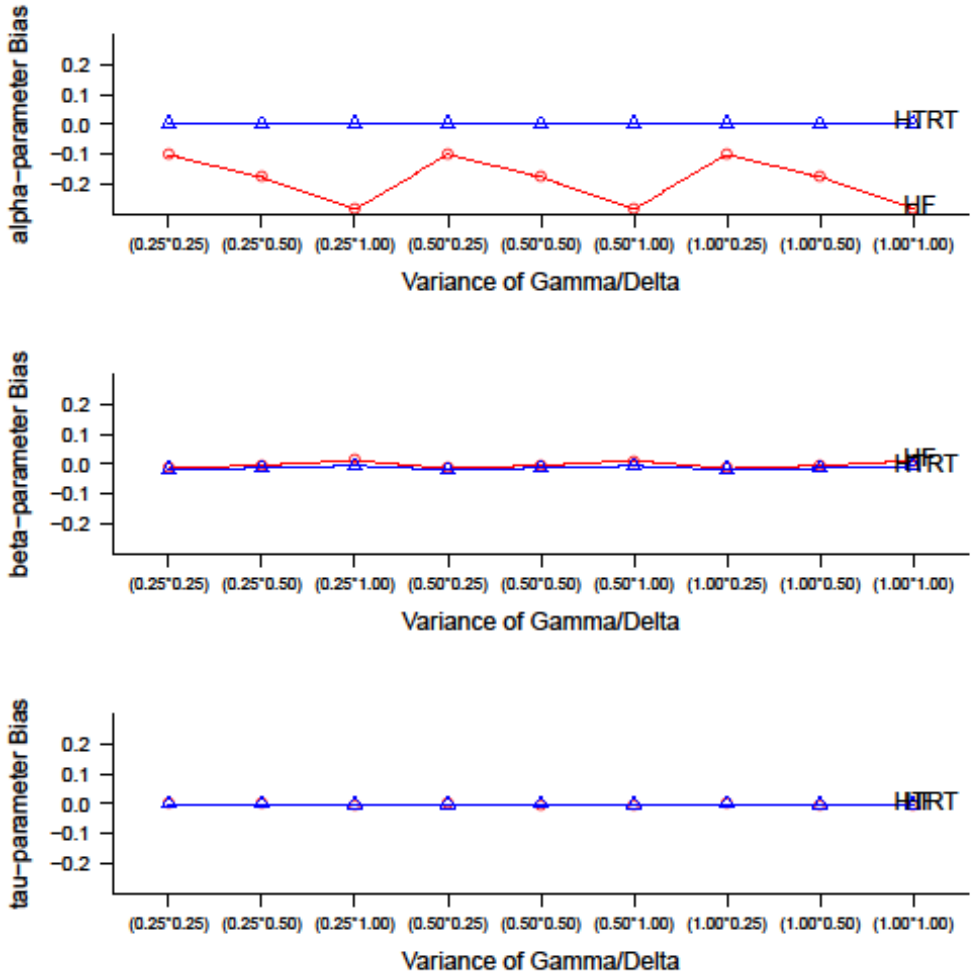

Figure 17. *Marginal bias in the recovery of response time parameters for grade 5. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

Tables 21 and 22 and Figures 18–23 present the results for marginal MSE (a measure of total error variability) of item and person parameters for both models. These MSE values were calculated by averaging the MSE within a replication, and calculating the mean and standard deviation of MSE over the replications. Most results were consistent with the bias results, with a few notable exceptions. Overall, the HTRT model showed lower MSE in all of the conditions and parameter estimation using the HTRT model was not affected by the presence of testlet variances. However, the MSE value for the time discrimination (α) parameter was not as high as the item discrimination (a) parameter. Even though the HTRT model had lower MSE for the ability (θ) and speed (τ) parameters than the Hierarchical Framework model, these parameters had relatively higher amounts of MSE when compared with their bias values.

Table 21

Mean MSE for response parameter estimates

		HTRT			Hierarchical Framework		
Grade	Condition	*a*	*b*	*θ*	*a*	*b*	*θ*
3	0.25 & 0.25	0.05	0.04	0.16	0.08	0.05	0.17
	0.25 & 0.50	0.05	0.04	0.16	0.08	0.05	0.17
	0.25 & 1.00	0.05	0.05	0.16	0.07	0.05	0.17
	0.50 & 0.25	0.05	0.05	0.19	0.12	0.06	0.21
	0.50 & 0.50	0.05	0.05	0.19	0.13	0.06	0.20
	0.50 & 1.00	0.06	0.05	0.19	0.13	0.06	0.20
	1.00 & 0.25	0.05	0.04	0.23	0.23	0.09	0.27
	1.00 & 0.50	0.05	0.04	0.23	0.24	0.09	0.27
	1.00 & 1.00	0.05	0.04	0.23	0.23	0.09	0.27
4	0.25 & 0.25	0.07	0.04	0.14	0.09	0.04	0.14
	0.25 & 0.50	0.07	0.04	0.14	0.09	0.04	0.14
	0.25 & 1.00	0.08	0.04	0.14	0.09	0.04	0.14
	0.50 & 0.25	0.07	0.04	0.16	0.14	0.04	0.17
	0.50 & 0.50	0.07	0.04	0.16	0.14	0.04	0.17
	0.50 & 1.00	0.07	0.04	0.16	0.13	0.04	0.17
	1.00 & 0.25	0.06	0.03	0.19	0.22	0.05	0.21
	1.00 & 0.50	0.06	0.03	0.19	0.22	0.05	0.21
	1.00 & 1.00	0.06	0.03	0.19	0.22	0.05	0.21
5	0.25 & 0.25	0.11	0.06	0.16	0.11	0.07	0.16
	0.25 & 0.50	0.11	0.06	0.16	0.11	0.07	0.16
	0.25 & 1.00	0.11	0.06	0.16	0.11	0.07	0.16
	0.50 & 0.25	0.11	0.06	0.17	0.13	0.07	0.18
	0.50 & 0.50	0.10	0.06	0.17	0.13	0.07	0.18
	0.50 & 1.00	0.11	0.06	0.17	0.14	0.07	0.18
	1.00 & 0.25	0.09	0.05	0.20	0.23	0.09	0.23
	1.00 & 0.50	0.09	0.06	0.20	0.23	0.09	0.22
	1.00 & 1.00	0.10	0.05	0.20	0.23	0.09	0.22

Table 22

Mean MSE for response time parameter estimates

		HTRT			Hierarchical Framework		
Grade	Condition	α	β	τ	α	β	τ
3	0.25 & 0.25	0.00	0.00	0.06	0.01	0.00	0.06
	0.25 & 0.50	0.00	0.00	0.09	0.03	0.00	0.12
	0.25 & 1.00	0.00	0.00	0.16	0.08	0.00	0.22
	0.50 & 0.25	0.00	0.00	0.06	0.01	0.00	0.06
	0.50 & 0.50	0.00	0.00	0.09	0.03	0.00	0.12
	0.50 & 1.00	0.00	0.00	0.16	0.08	0.00	0.22
	1.00 & 0.25	0.00	0.00	0.06	0.01	0.00	0.06
	1.00 & 0.50	0.00	0.00	0.09	0.03	0.00	0.12
	1.00 & 1.00	0.00	0.00	0.16	0.08	0.00	0.22
4	0.25 & 0.25	0.00	0.00	0.03	0.01	0.00	0.04
	0.25 & 0.50	0.00	0.00	0.05	0.03	0.00	0.07
	0.25 & 1.00	0.00	0.00	0.06	0.08	0.00	0.12
	0.50 & 0.25	0.00	0.00	0.03	0.01	0.00	0.04
	0.50 & 0.50	0.00	0.00	0.05	0.03	0.00	0.07
	0.50 & 1.00	0.00	0.00	0.06	0.08	0.00	0.12
	1.00 & 0.25	0.00	0.00	0.03	0.01	0.00	0.04
	1.00 & 0.50	0.00	0.00	0.05	0.03	0.00	0.07
	1.00 & 1.00	0.00	0.00	0.06	0.08	0.00	0.12
5	0.25 & 0.25	0.00	0.00	0.03	0.01	0.00	0.03
	0.25 & 0.50	0.00	0.00	0.04	0.04	0.00	0.06
	0.25 & 1.00	0.00	0.00	0.05	0.10	0.00	0.10
	0.50 & 0.25	0.00	0.00	0.03	0.01	0.00	0.03
	0.50 & 0.50	0.00	0.00	0.04	0.04	0.00	0.06
	0.50 & 1.00	0.00	0.00	0.05	0.10	0.00	0.10
	1.00 & 0.25	0.00	0.00	0.03	0.01	0.00	0.03
	1.00 & 0.50	0.00	0.00	0.04	0.04	0.00	0.06
	1.00 & 1.00	0.00	0.00	0.05	0.10	0.00	0.10

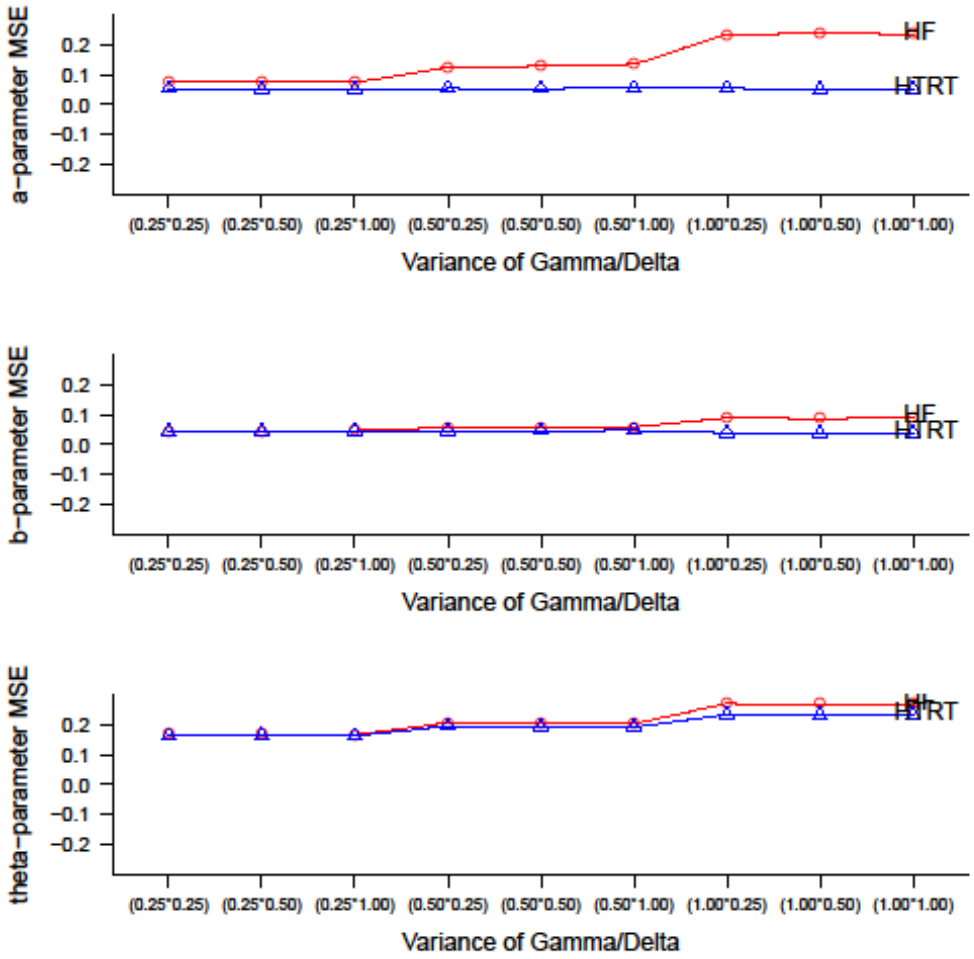

Figure 18. *Marginal MSE in the recovery of response parameters for grade 3. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

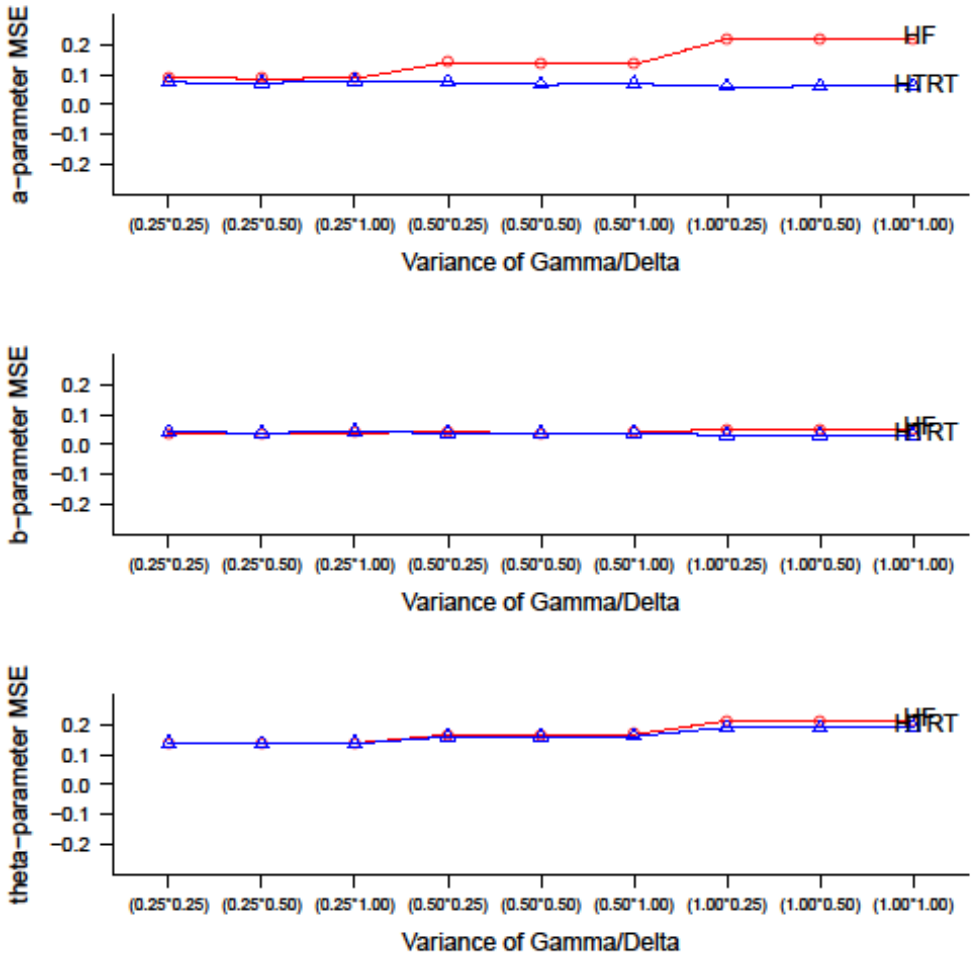

Figure 19. *Marginal MSE in the recovery of response parameters for grade 4. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

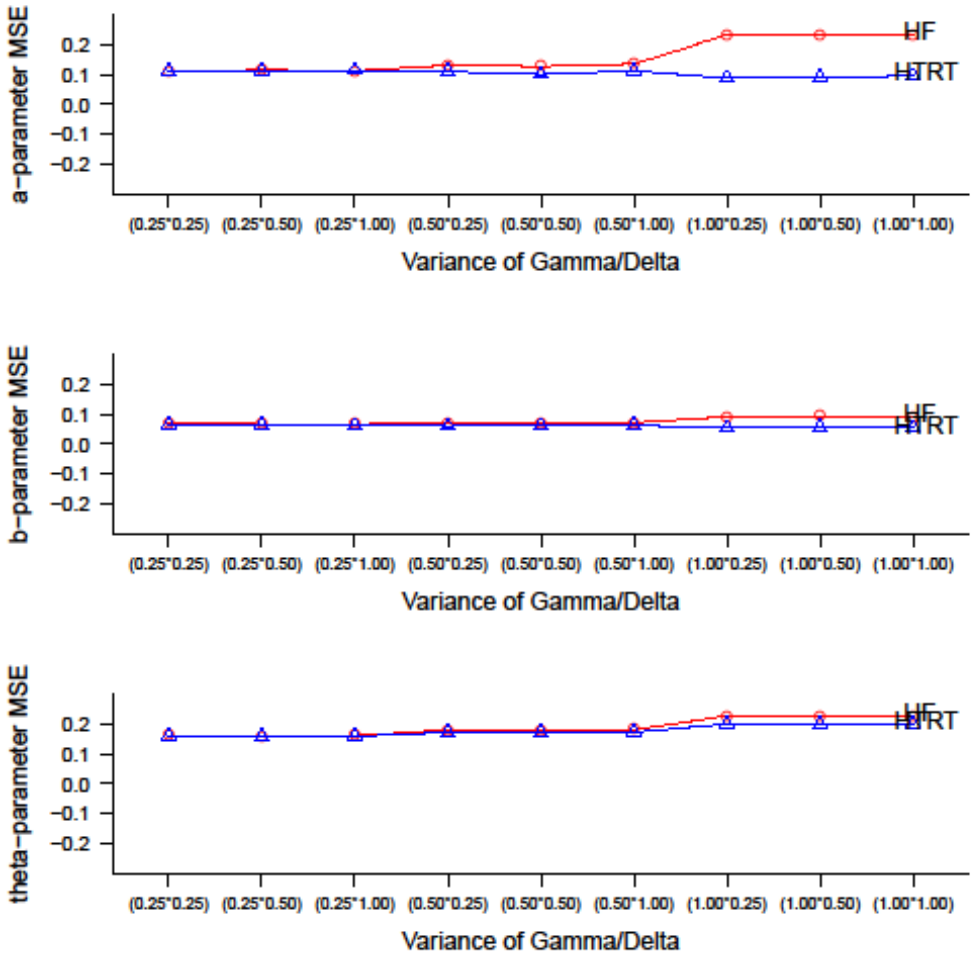

Figure 20. *Marginal MSE in the recovery of response parameters for grade 5. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

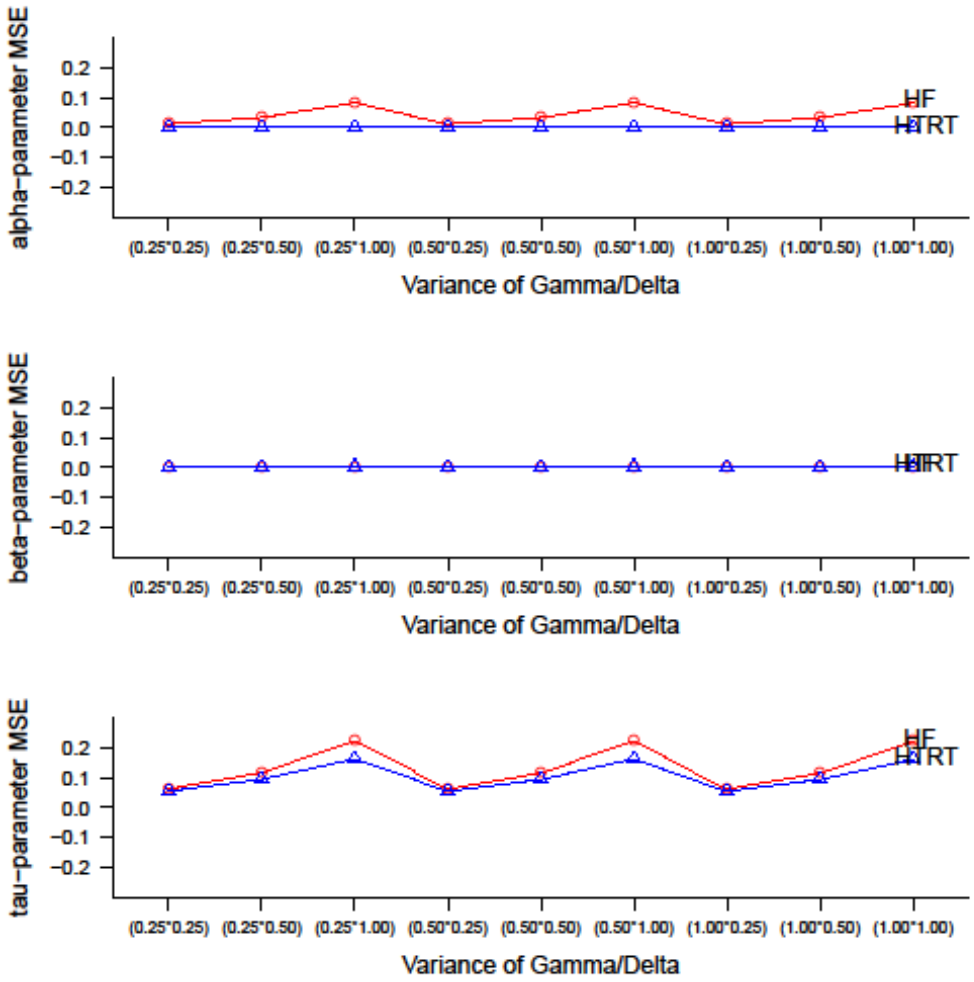

Figure 21. *Marginal MSE in the recovery of response time parameters for grade 3. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

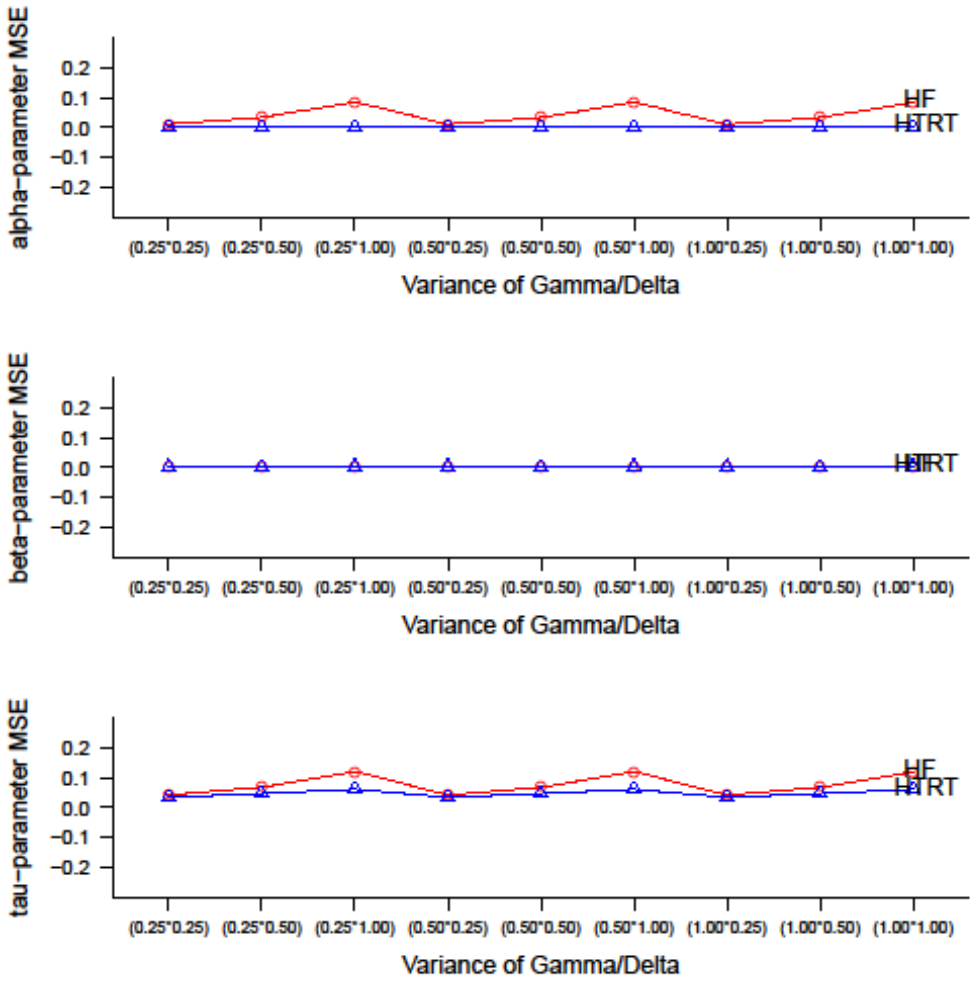

Figure 22. *Marginal MSE in the recovery of response time parameters for grade 4. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

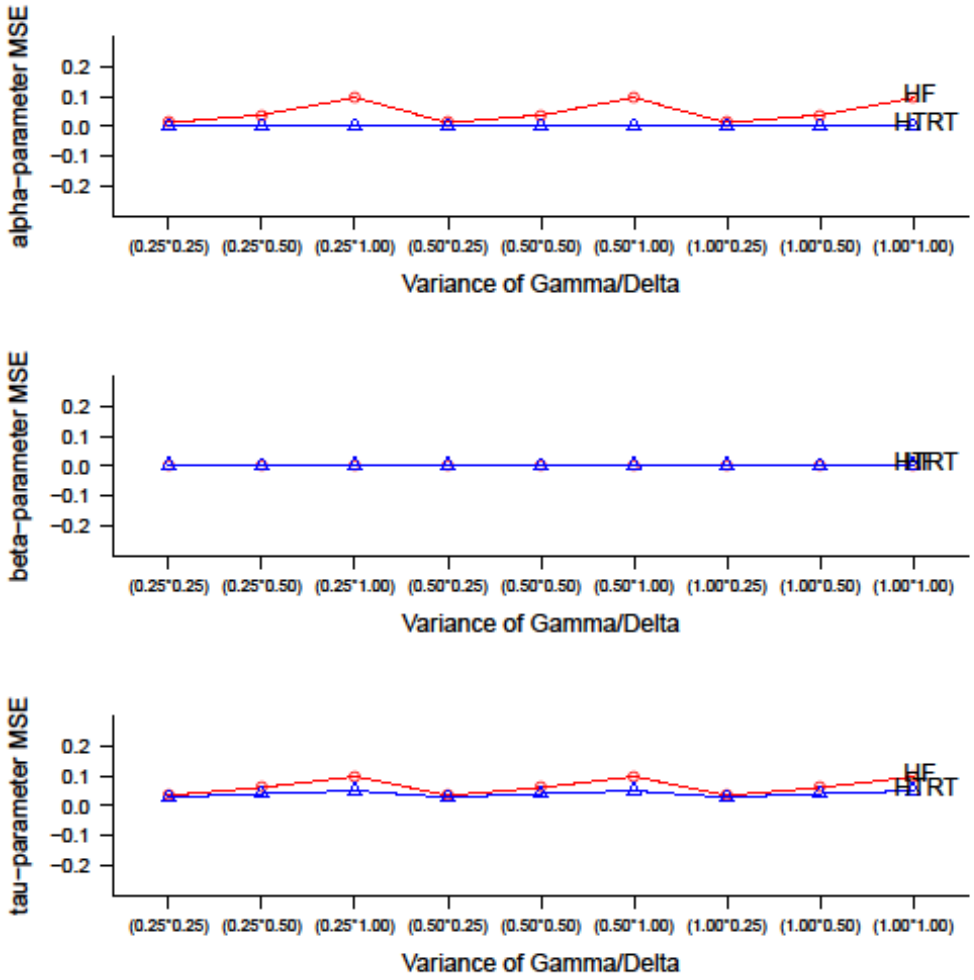

Figure 23. *Marginal MSE in the recovery of response time parameters for grade 5. Lines are labelled to correspond models (HTRT=Hierarchical Testlet Response Time (blue) & HF=Hierarchical Framework (red)).*

The scatter plots of the true and estimated parameters for all three grades of both models are presented in Figures A10–A27 in the Appendix. The scatter plots show similar patterns of results to those of the bias and MSE for both models. The points on the scatter plots for the HTRT model are generally located near to the reference line (a 45 degree line through the origin), which indicates smaller errors in parameter estimation.

TCC and TIF

The TCC for both models were created with response (a, b, and θ) and response time (α, β, and τ) parameters and shown in Figures 24–25. The inflection point of the TCC for response parameters is at the low end of the ability distribution, while that of the response time TCC is at the high end of the speed distribution. The inflections of the curve for both response and response time parameters were similar to those observed with real data. The examinees with the average ability were expected to have high scores but the examinees with the average speed were expected to have very low total scores. The assessments were easy enough for examinees with lower abilities to have high expected scores, and examinees who were fast and took a small amount of testing time had high expected scores.

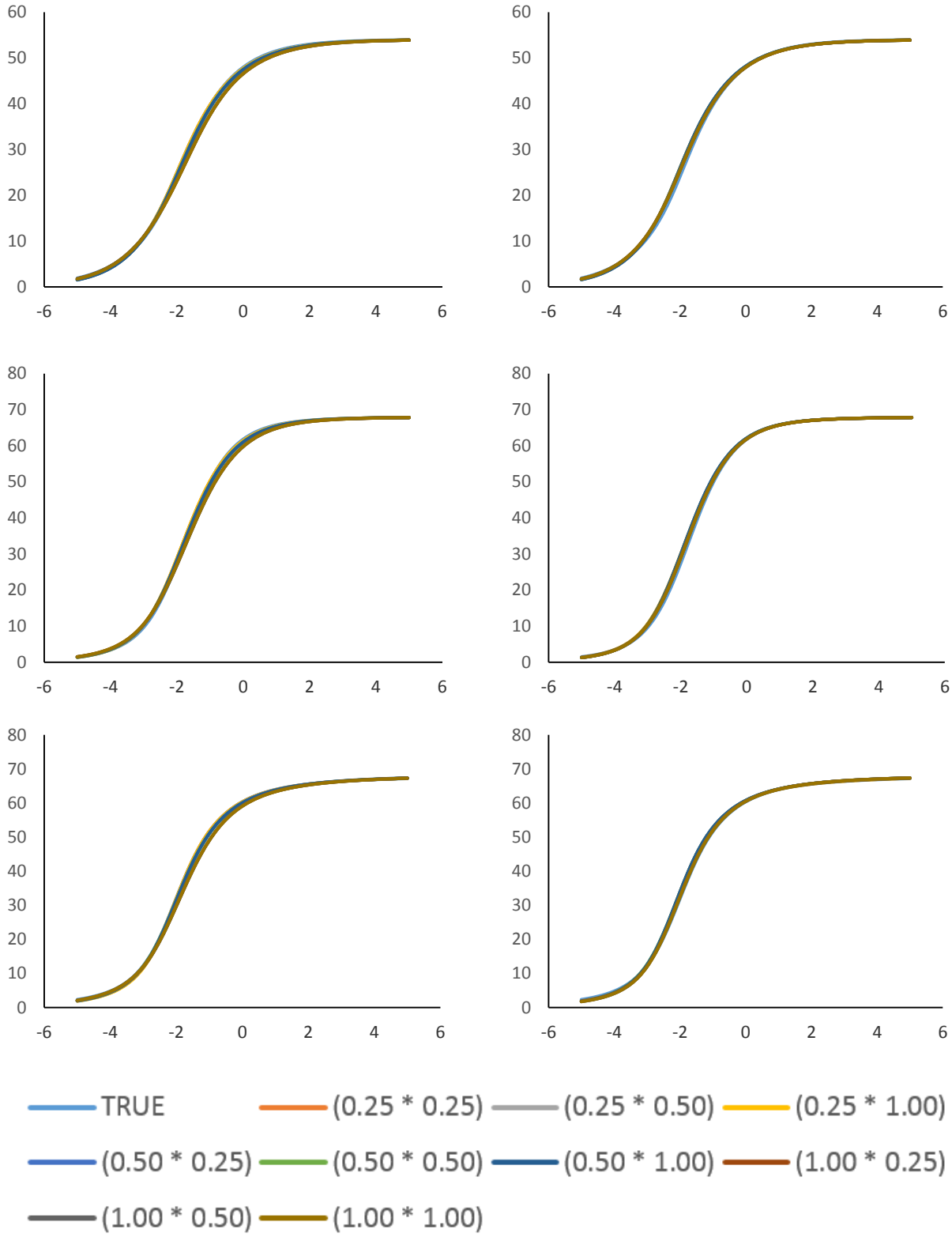

Figure 24. *Test characteristic curves using response parameters of the HTRT model (right) and the Hierarchical Framework model (left) for grade 3 (top), grade 4 (center), and grade 5 (bottom).*

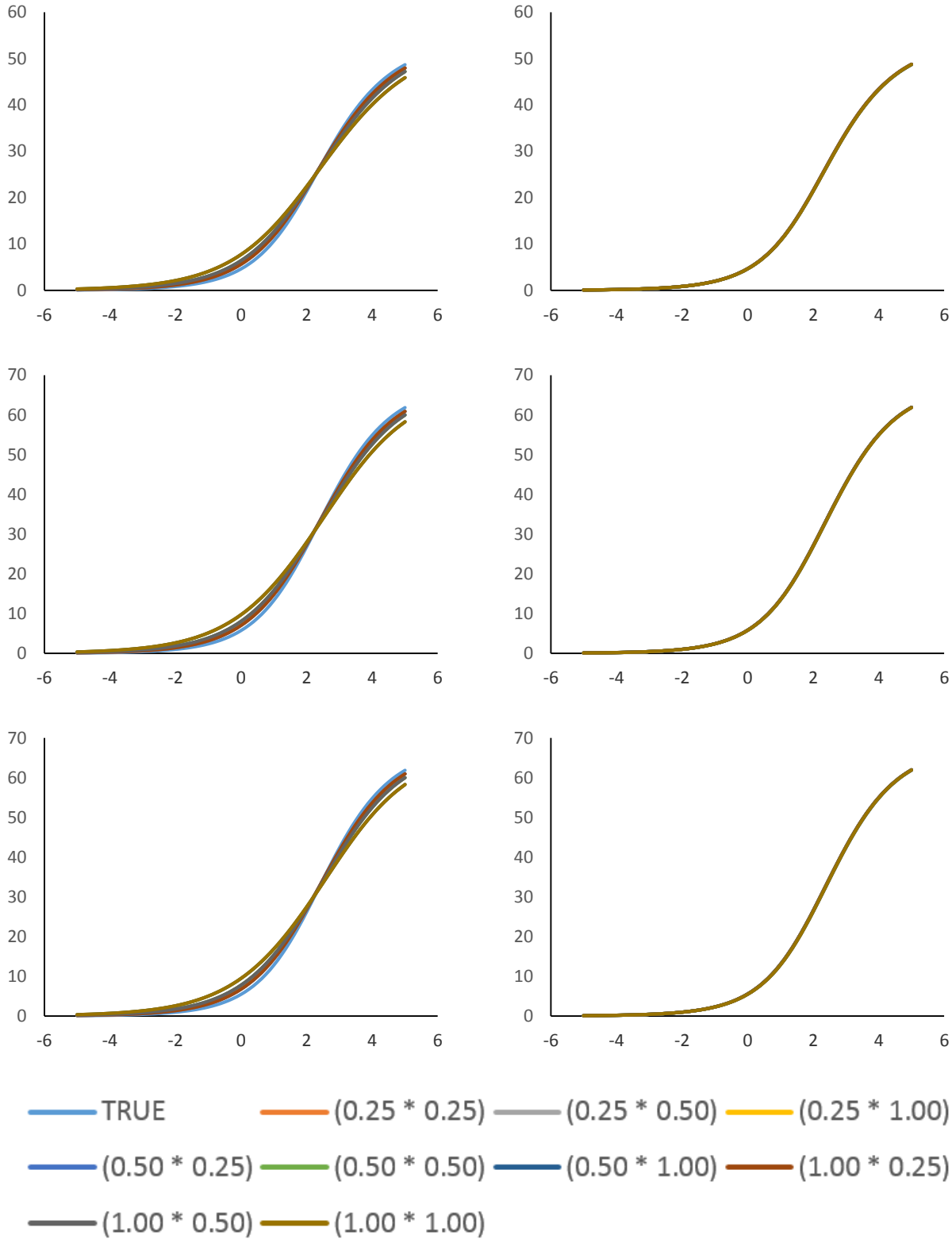

Figure 25. *Test characteristic curves using response time parameters of the HTRT model (right) and the Hierarchical Framework model (left) for grade 3 (top), grade 4 (center), and grade 5 (bottom).*

The TIF for both models are shown in Figures 26–31. Test information for all three grades was maximized around ability (θ) values of -2.0. For response time parameters, test information for all three grades was maximized around speed (τ) values of 2.0. For the HTRT model, the estimated TIFs for all nine conditions were very similar to the true TIFs. For the Hierarchical Framework model, the underestimation of test information was dramatic as the amount of testlet variances increased. Thus, fitting a unidimensional model (i.e., Hierarchical Framework) produced seriously flawed estimates of score reliability when the assumption of local independence was violated (even though the models all converged).

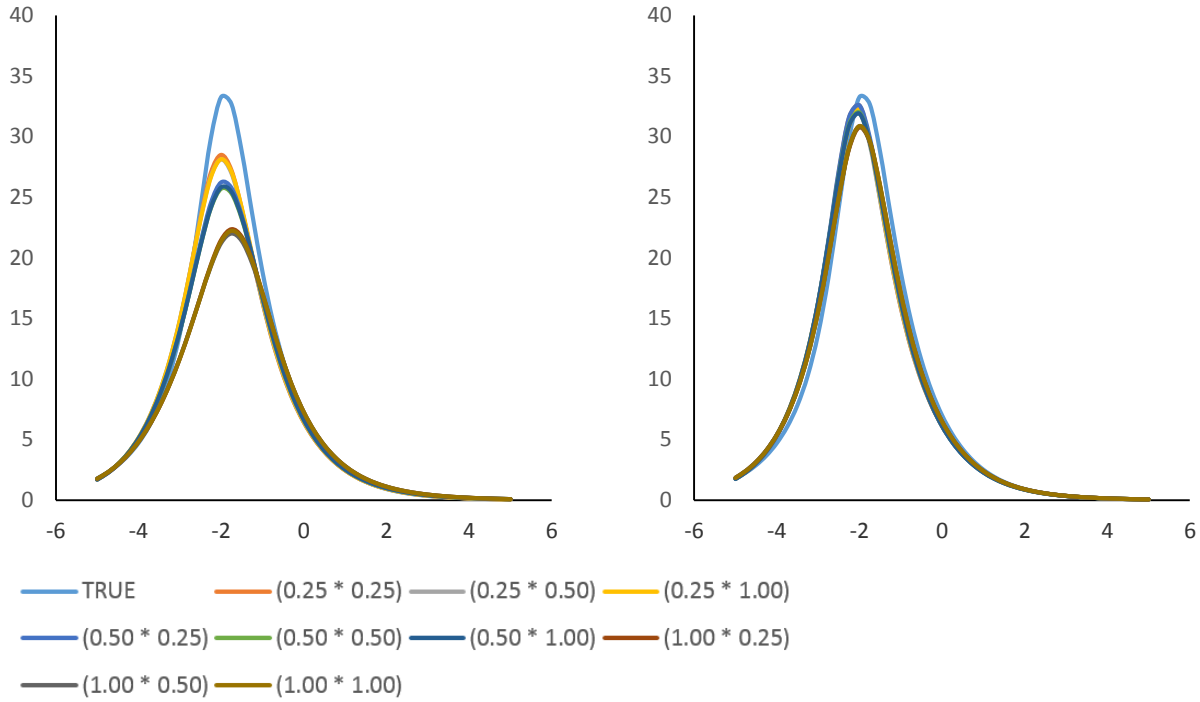

Figure 26. *Test information function using response parameters for the HTRT model (right) and the Hierarchical Framework model (left) with nine testlet conditions of grade 3.*

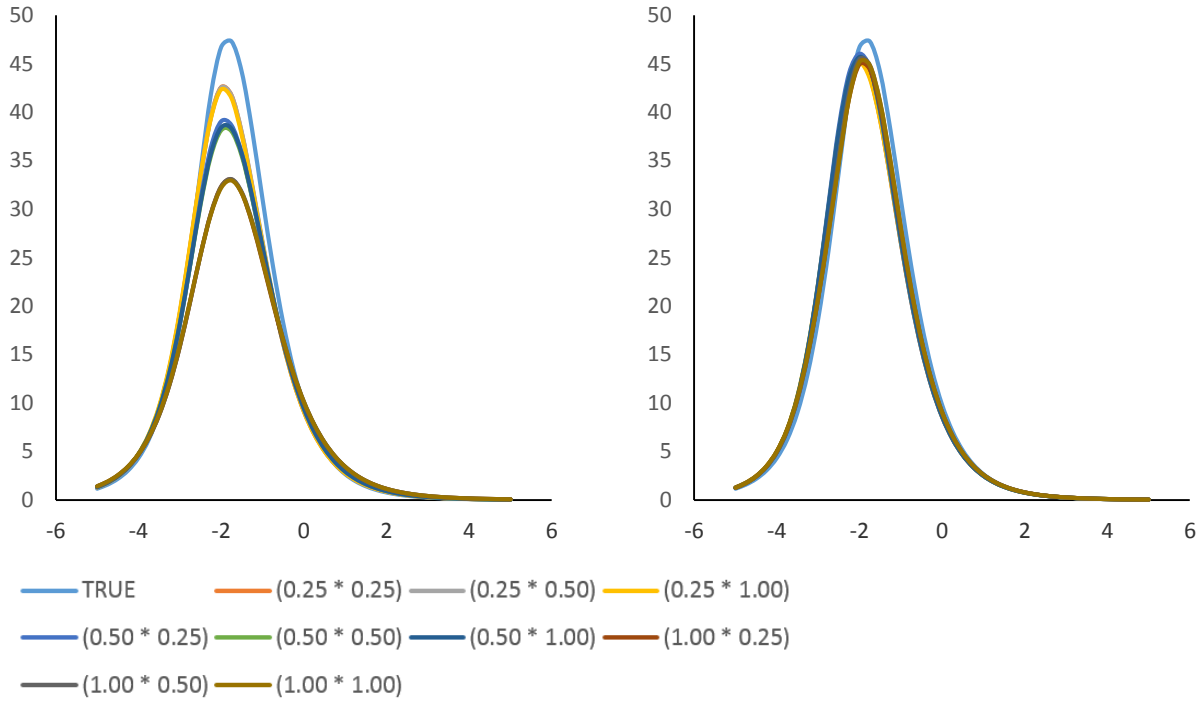

Figure 27. *Test information function using response parameters for the HTRT model (right) and the Hierarchical Framework model (left) with nine testlet conditions of grade 4.*

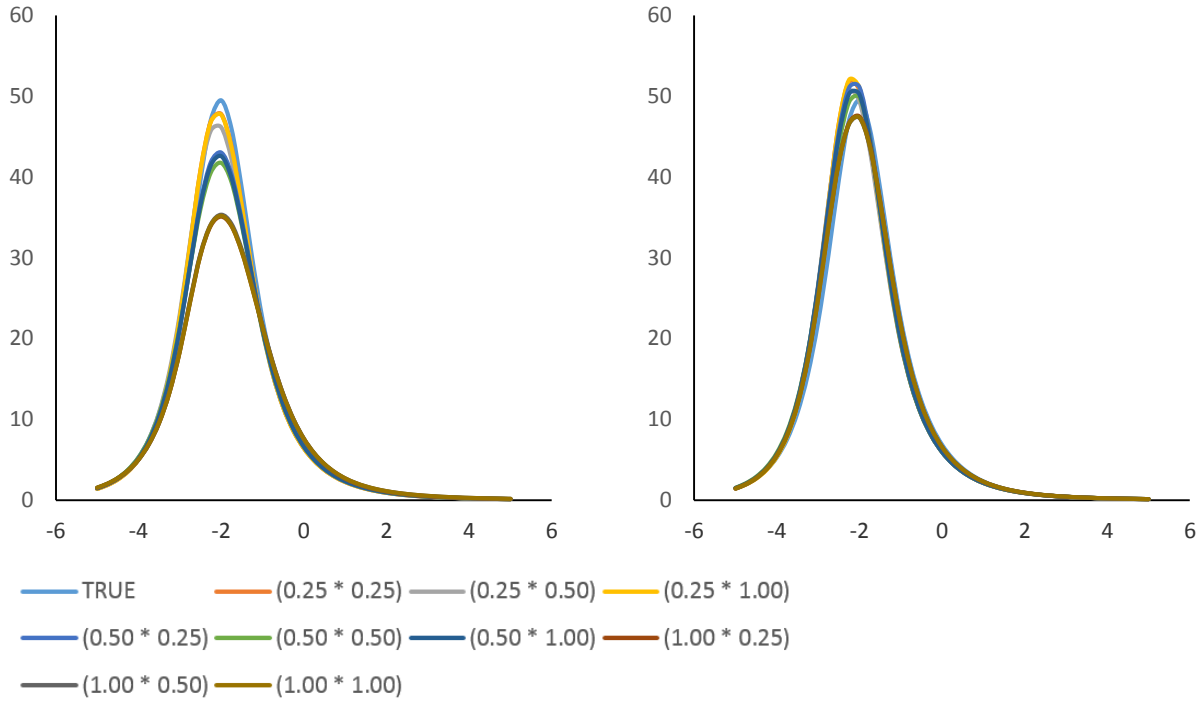

Figure 28. *Test information function using response parameters for the HTRT model (right) and the Hierarchical Framework model (left) with nine testlet conditions of grade 5.*

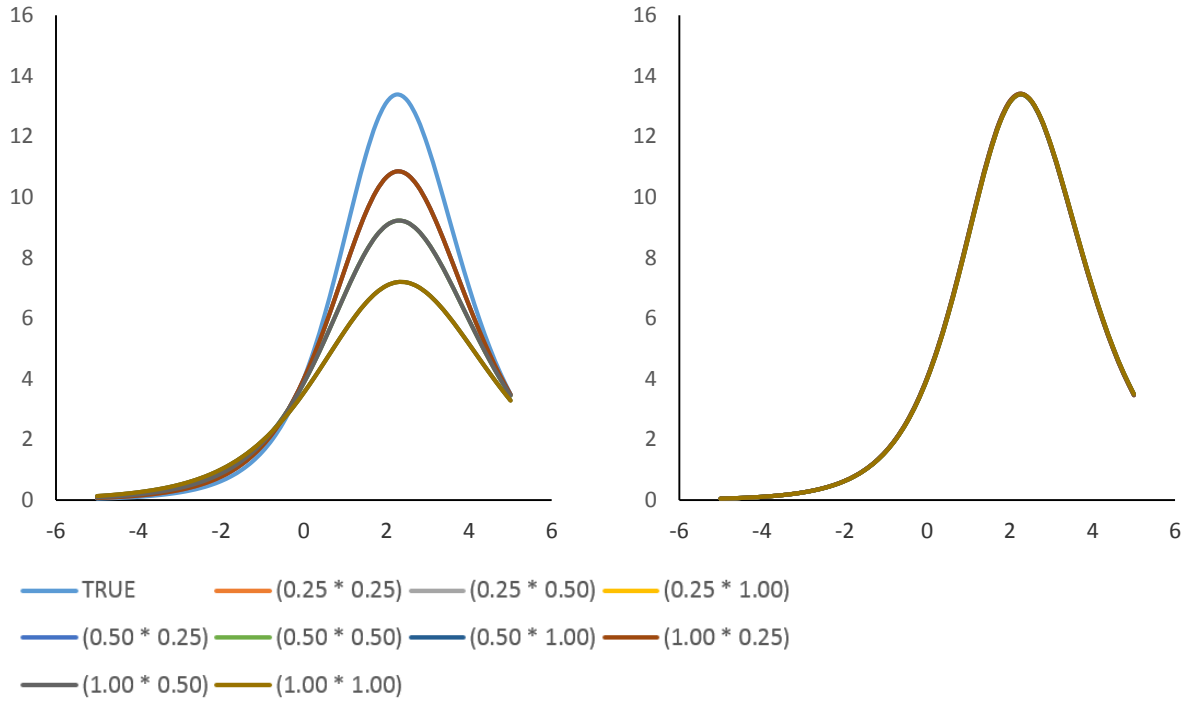

Figure 29. *Test information function using response time parameters for the HTRT model (right) and the Hierarchical Framework model (left) with nine testlet conditions of grade 3.*

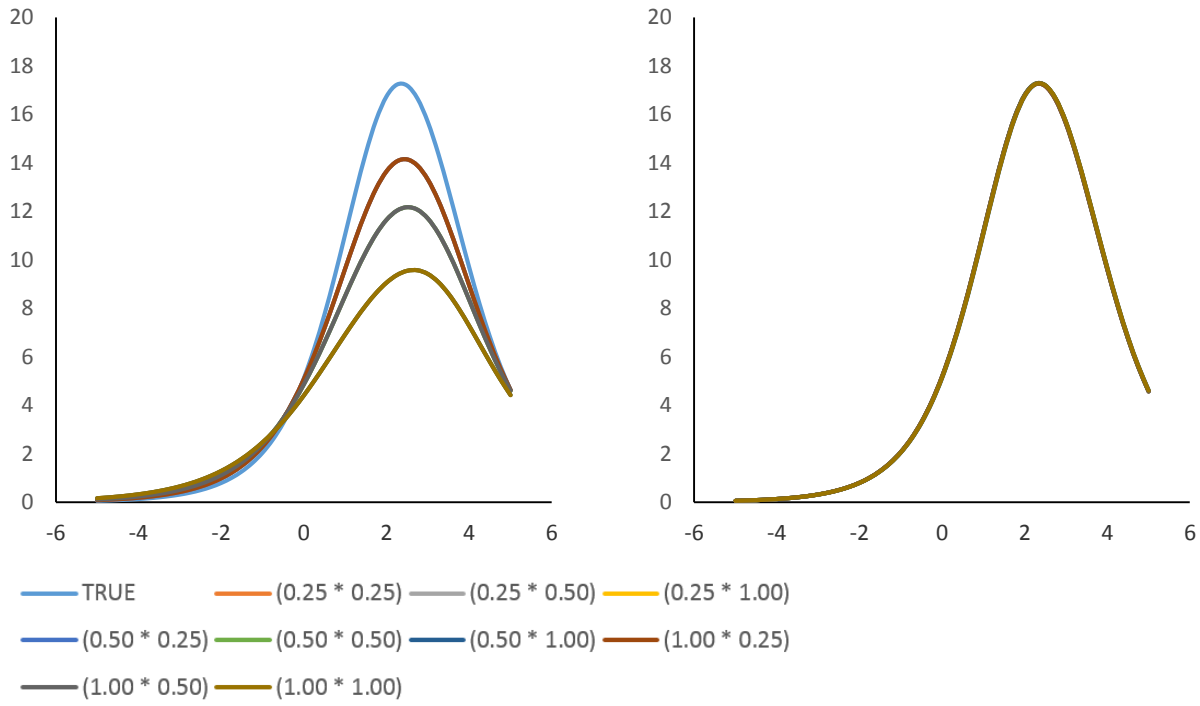

Figure 30. *Test information function using response time parameters for the HTRT model (right) and the Hierarchical Framework model (left) with nine testlet conditions of grade 4.*

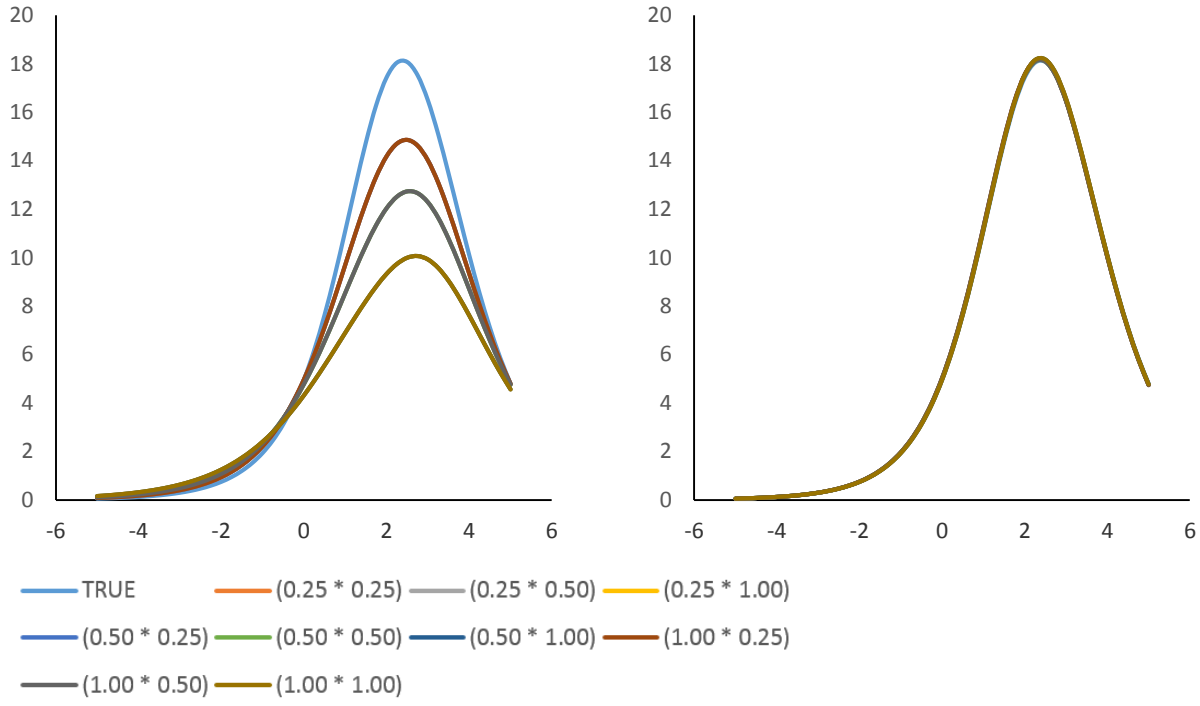

Figure 31. *Test information function using response time parameters for the HTRT model (right) and the Hierarchical Framework model (left) with nine testlet conditions of grade 5.*

Conditional Theta and Speed with Bias and MSE

The examinee parameter recovery was further examined by categorizing examinees. Twelve groups were created based on examinee true ability and speed, and each group's average bias and MSE of estimated ability and speed were calculated.

Figures 32–34 present examinees' conditional bias and MSE in regards to their true ability levels. For the conditional bias, ability was overestimated for lower-ability examinees and underestimated for higher-ability examinees. MSE was higher for examinees with higher ability levels. In general, the conditions with higher γ variances had higher amounts of bias and MSE. Patterns for both models were similar.

Figures 35– 37 present examinees' conditional bias and MSE in regards to their true speed levels. Bias and MSE for the speed parameter were lower when compared with the ability parameter. The bias and MSE values were more constant with regard to the true values when compared to the ability parameter. However, the bias in speed parameter estimates for grade 3 was dissimilar to the rest of the grades but more closely resembled patterns for the bias in ability parameter. Generally, the conditions with higher variance in the δ-parameter had higher amounts of bias and MSE.

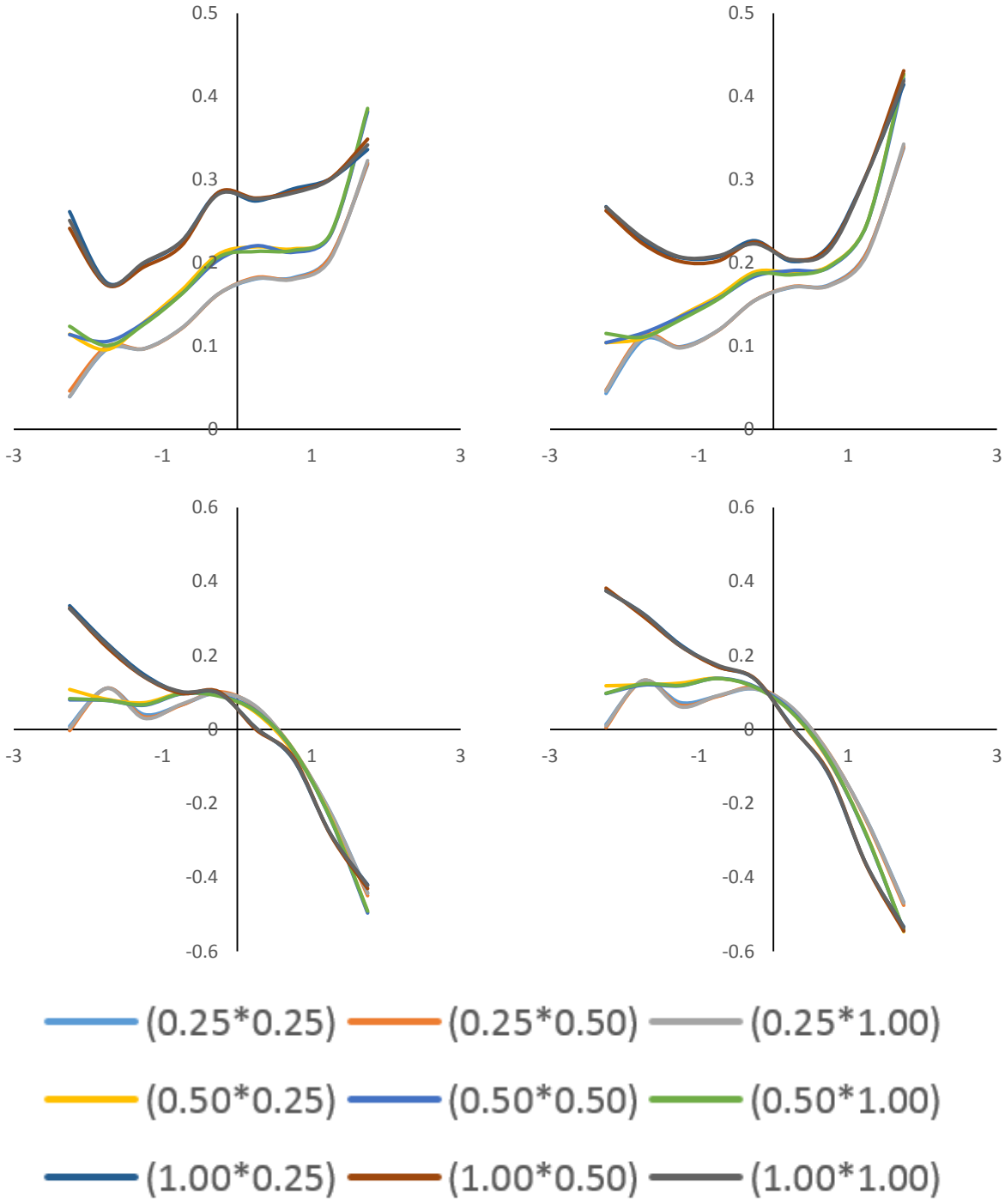

Figure 32. *Conditional MSE (top) and bias (bottom) across theta of the HTRT model (right) and the Hierarchical Framework model (left) for grade 3.*

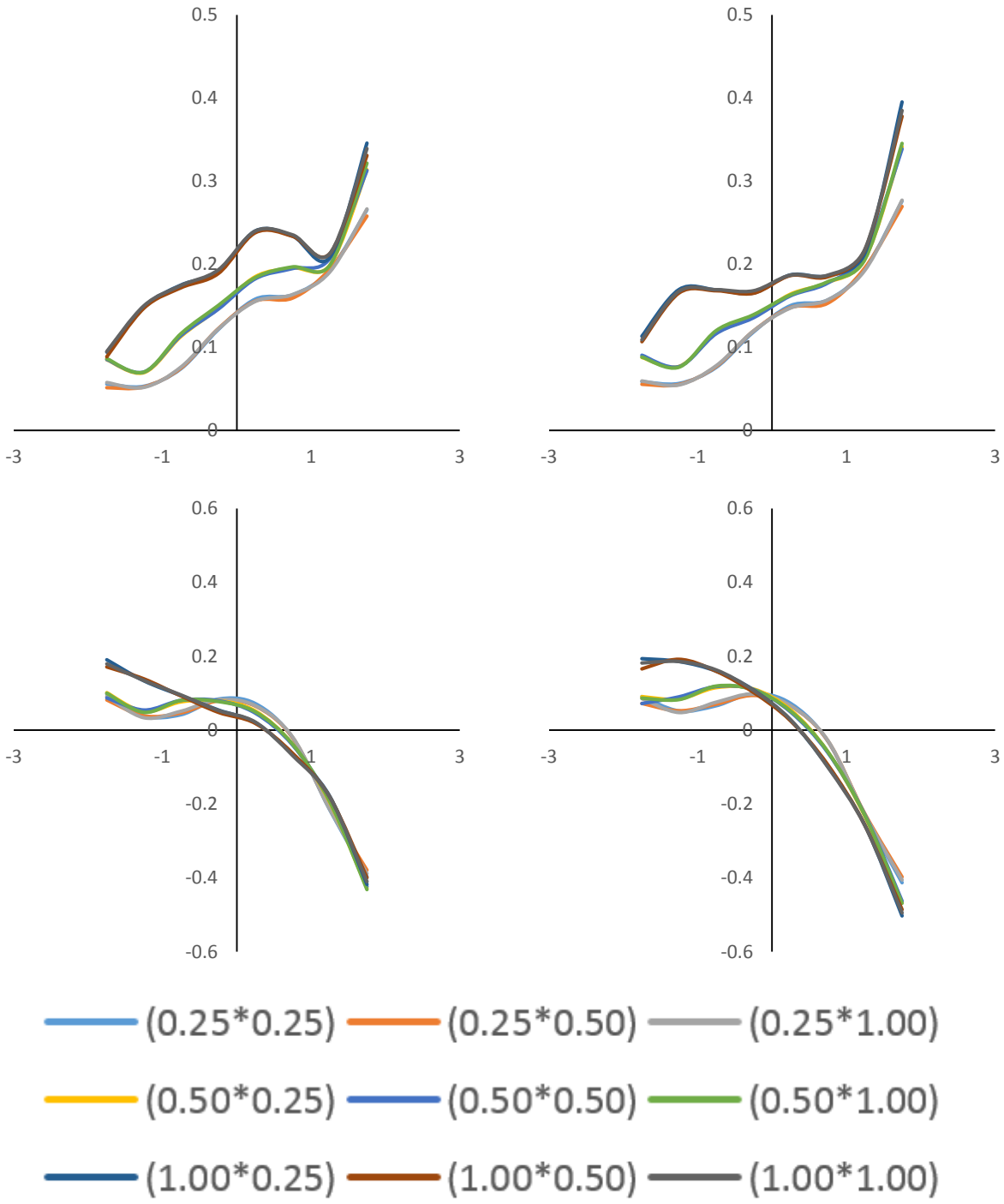

Figure 33. *Conditional MSE (top) and bias (bottom) across theta of the HTRT model (right) and the Hierarchical Framework model (left) for grade 4.*

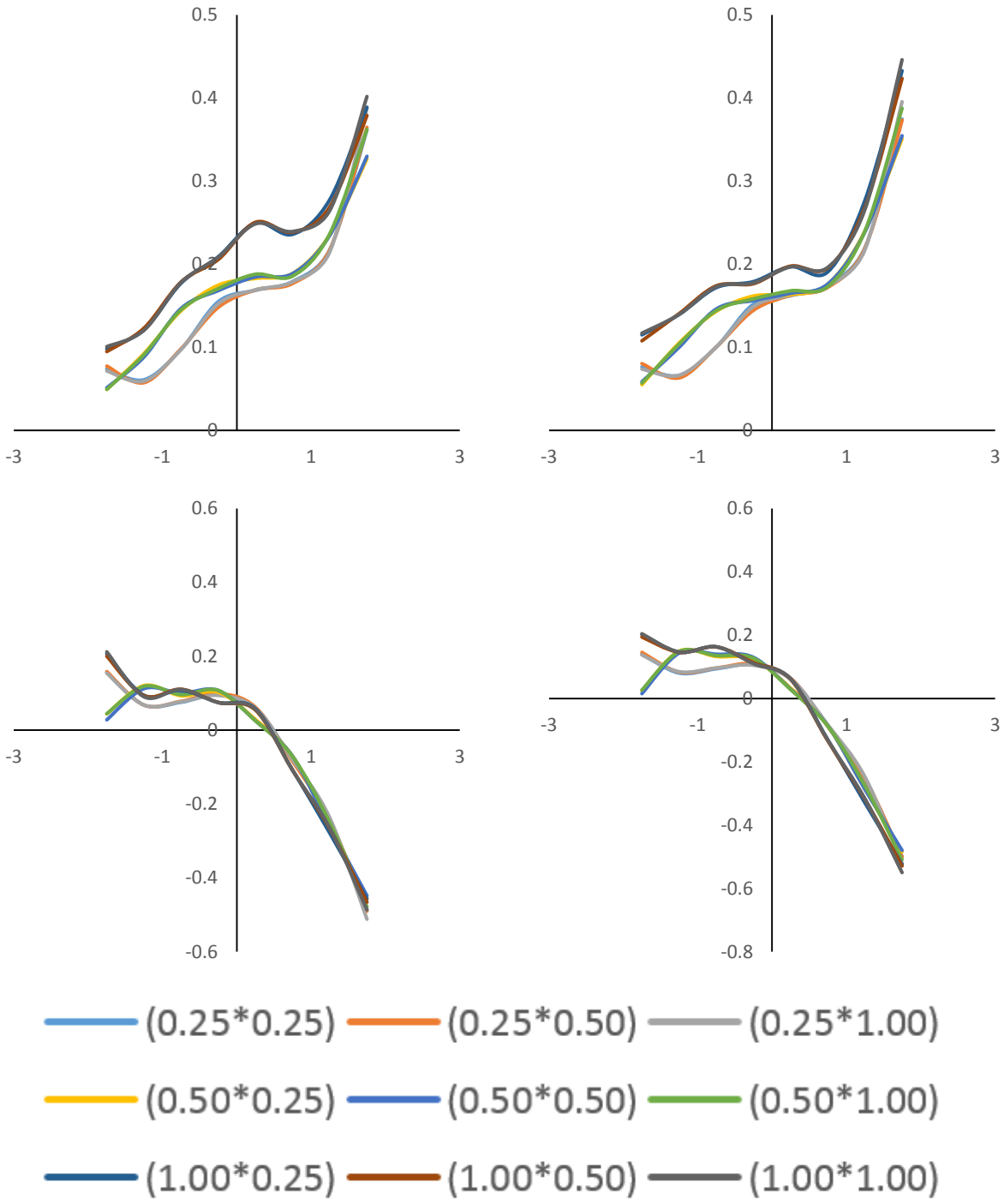

Figure 34. *Conditional MSE (top) and bias (bottom) across theta of the HTRT model (right) and the Hierarchical Framework model (left) for grade 5.*

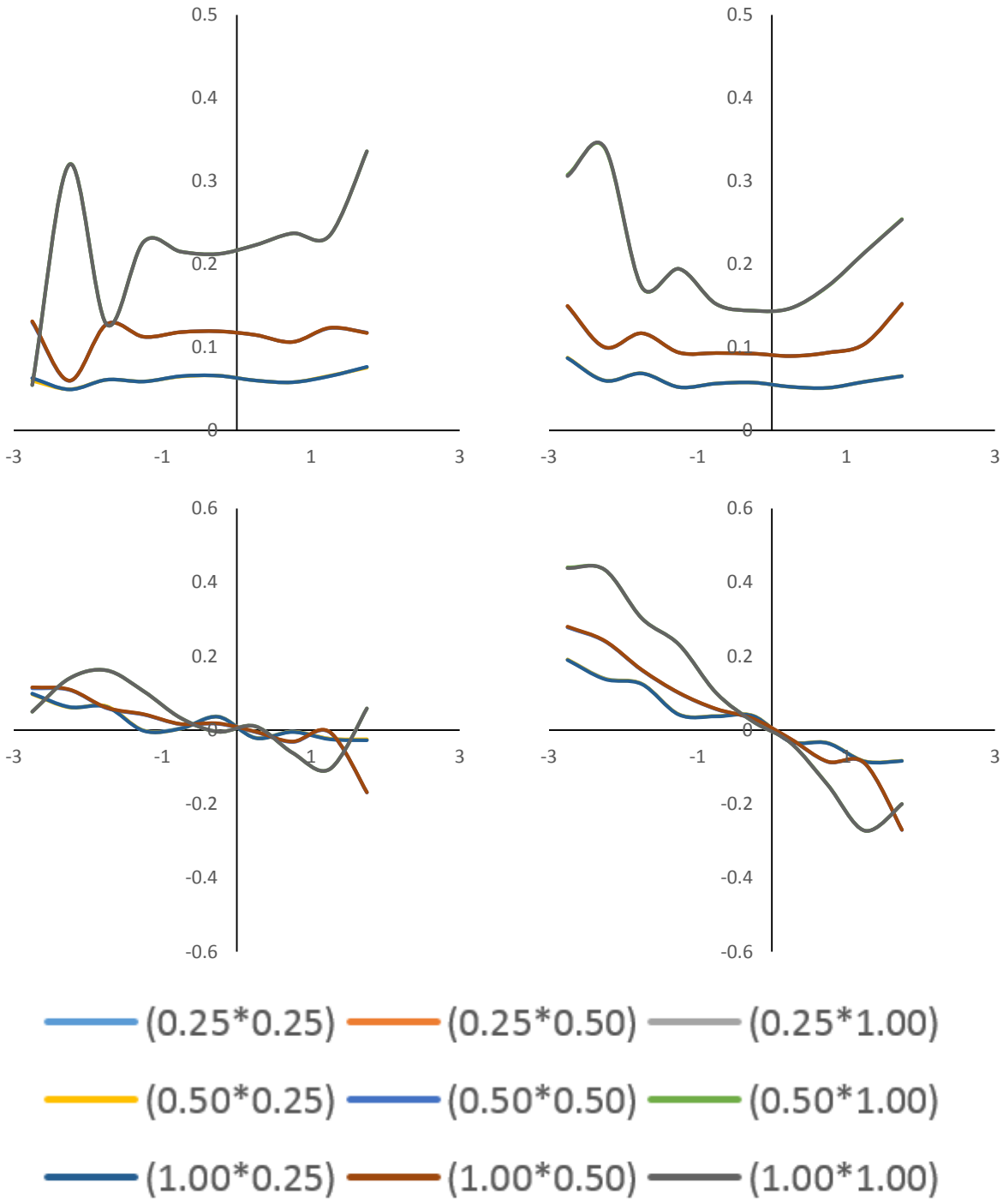

Figure 35. *Conditional MSE (top) and bias (bottom) across tau of the HTRT model (right) and the Hierarchical Framework model (left) for grade 3.*

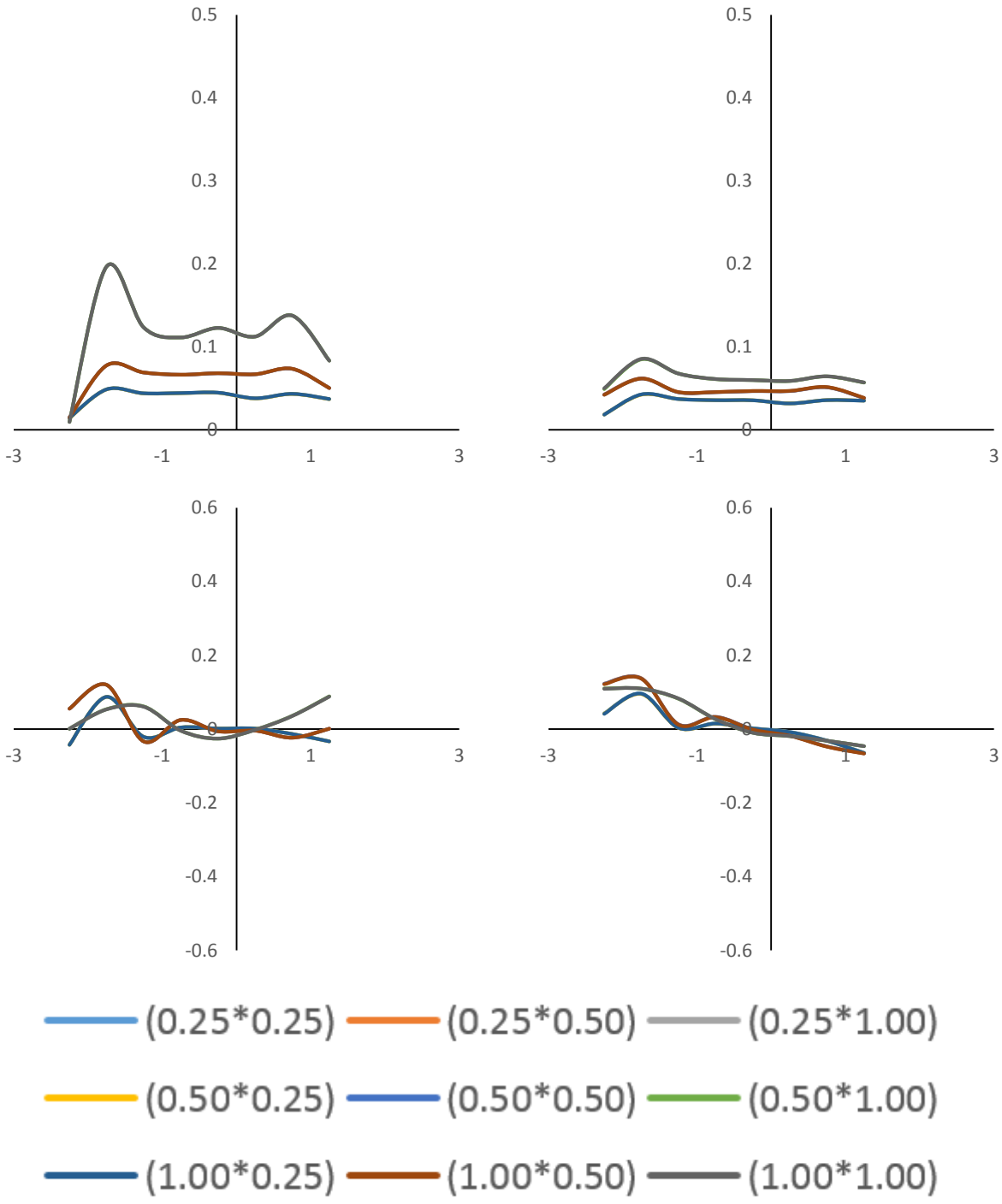

Figure 36. *Conditional MSE (top) and bias (bottom) across tau of the HTRT model (right) and the Hierarchical Framework model (left) for grade 4.*

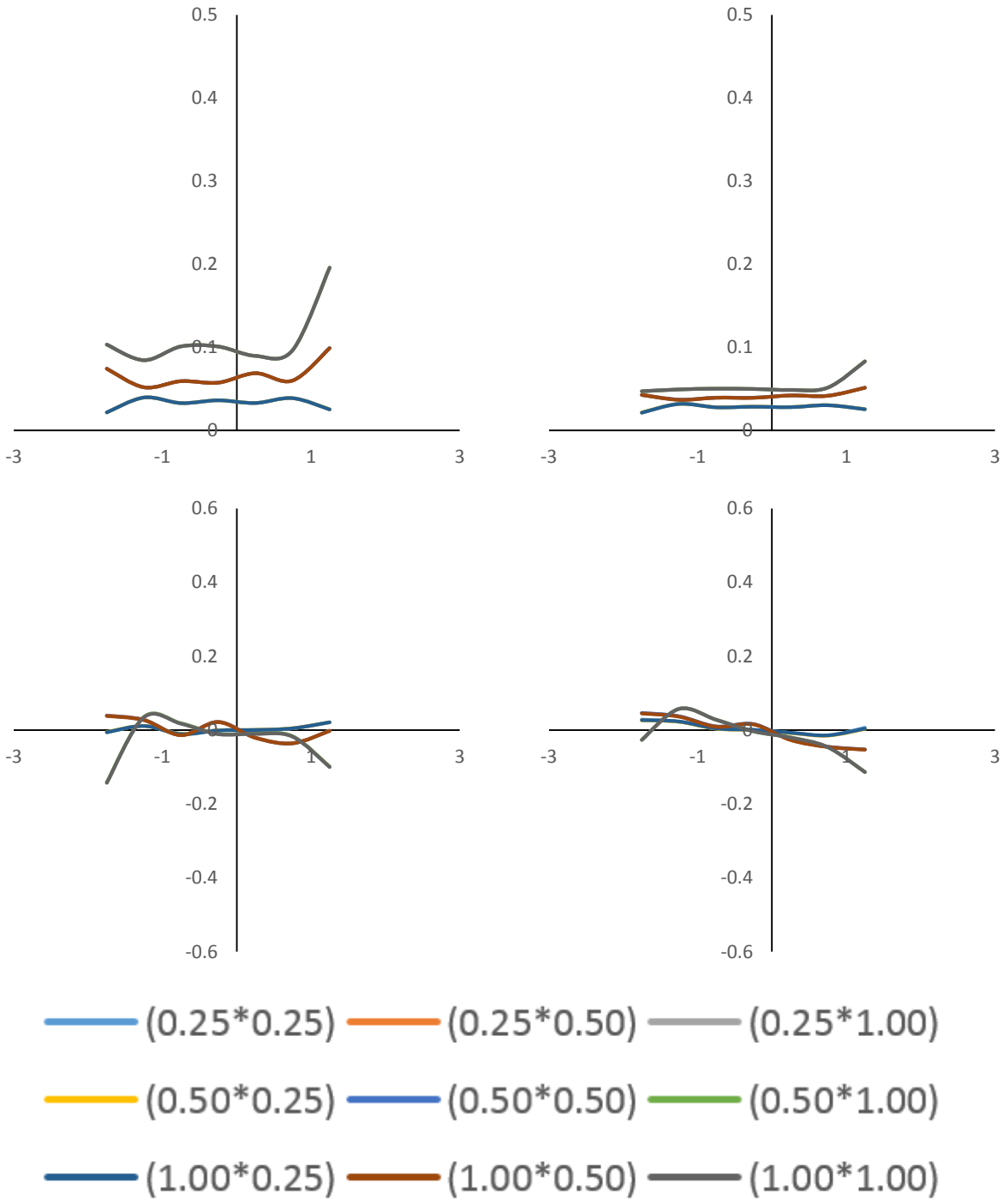

Figure 37. *Conditional MSE (top) and bias (bottom) across tau of the HTRT model (right) and the Hierarchical Framework model (left) for grade 5.*

Chapter 5. Discussion

The purpose of this study was to introduce a new scoring response time model using response and response time data to address testlet effects. Two response time models (HTRT & Hierarchical Framework) were compared, and the parameter recovery was examined using real and simulated data. The research questions addressed in this study are as follows:

1. If the local independence assumption is violated, then how much improvement does the HTRT model provide over the Hierarchical Framework model (van der Linden, 2007) in parameter estimation?
2. How do various test conditions impact parameter estimation and possibly cause estimation errors?

These research questions were addressed using the overall results from the comparison of the response time models (HTRT and Hierarchical Framework) when the local independence assumption was violated. The discussion of these topics is followed by a listing of the limitations of the study. As a conclusion, the implications of this work for educational practice and further research are presented.

Study 1

Study 1 applied the HTRT model to real data in order to obtain parameter estimates and the MCMC component specifications for Study 2. The assessment for all three grades showed good to excellent internal consistency for responses and response times.

The HTRT model was implemented using the *R2OpenBUGS* package (Sturtz, Ligges, & Gelman, 2010) from R (R Core Team, 2014). The final length for the Markov Chain was 15,000 and the burn-in period was 10,000. The last 5,000 draws from two separate Markov Chains were retained for inference making. The data for all three grades were assumed to be converged.

Overall, items had high discrimination and lower difficulty, and examinees took a great amount of time on items. The estimated parameters had very little or no relationship between them. The only exception was the correlation between the time discrimination (α) and time intensity (β) parameters of grade 3, which had a moderate level of negative relationship.

The TCCs indicated that the assessments were very easy for examinees with lower abilities to have high expected scores, and examinees who were fast and took small amount of testing time had high expected scores. The TIFs showed that test information is maximized around ability (θ) values of -2.0. Test information was also maximized around speed (τ) values of 2.0. The amount of information was very low with response time parameters when compared to the information with response parameters.

Study 2

Overall, the parameter estimates using the HTRT model were steady and were not affected by the presence of testlet variances. Simply, this occurred because the HTRT model does include extra parameter to address the local dependency among testlet items. The parameter estimates using the Hierarchical Framework model were affected by the amount of shared variance among testlet items. The item discrimination and time discrimination parameters showed the most variation across the nine testlet conditions, followed by the item difficulty parameter.

The HTRT model showed lower marginal bias for item discrimination, item difficulty, and time discrimination parameters. The amount of marginal bias for the Hierarchical Framework model increased as the amount of variance for testlet parameters increased. The parameters from the response model were affected by the amount of shared variance of the γ-

parameter, and the parameters from response time model were affected by the amount of shared variance of δ-parameter. This was expected since the response model (e.g., 3-PL TRT) does include γ-parameter and the response time model (e.g., lognormal with testlet parameter) does include δ-parameter. The outcome definitely indicated that a significant amount of systematic error variance is presented for the item discrimination and time discrimination parameters, and slightly less amount of systematic error variance for the item difficulty parameter. The systematic error is not determined by chance but is introduced by applying incorrect model (Hierarchical Framework) to the data. Also, for incorrect model, the increase of shared variance within a testlet certainly had the amount of systematic error increased.

The findings for marginal MSE were very similar to the findings for marginal bias. Since the average MSE represents a total error variance, parameters with high systematic error variance also had high total error variance. One interesting point was that both the examinee ability and speed parameters had relatively higher amount of total error variance with nearly no systematic error variance. It is possible that greater amounts of random errors are affecting these two parameters and nearly no systematic error is involved. The random errors are unpredictable and their expected values are scattered around the true value. However, it is clear that random errors are also affected by the amount of shared variance within testlet items.

Based on the TIF results, the HTRT model indicated that each ability level is estimated very well with the data. The amount of information or precision is very similar to the true level. However, the Hierarchical Framework model showed lower information across the ability level. The incorrect model (Hierarchical Framework) and the presence of shared variance among testlet items did affect the estimation by producing less test information.

Limitations of the study and further research questions

There are several limitations and a number of issues for future studies. First, the test format was very limited for the simulation study because it followed the format of the real data. Varying the number of examinees, the test length, the number of testlet items, or the number of independent items could provide more information about how these test formats can affect the estimation. Second, this study excluded examinees who did not complete the assessment. The data for this study had very high average total scores and very low item difficulty parameters. It would be interesting to investigate the data including examinees who did not complete the assessment. Third, the DIC has a tendency to select the more complex model (Kang & Cohen, 2007; Li, Cohen, Kim & Cho, 2009). Additional model-fit indices that could be used to select the best model include Akaike's information criteria (AIC), Bayesian information criteria (BIC), pseudo Bayes factor (PsBF), posterior model checks (PPMC) and cross validation loglikelihood (CVLL). However, the application of these indices is usually done with item response models and the usage for response time models would require extensive research to determine whether they behave similarly to item response models. Fourth, this study considered only a single convergence criterion. It would be beneficial to use multiple convergence criteria to determine convergence. Finally, this study is based on the Hierarchical Framework model to address the testlet effect. Future studies with more response time models to investigate the testlet effect would provide very resourceful information.

Conclusion

The Bayesian estimation using the MCMC method was applied to compare the response time models. The HTRT model was introduced to address shared variance among testlet items. The HTRT model produced better parameter recovery than the Hierarchical Framework model.

Although this study may include several practical issues, there have been no response time models to address testlet effect. The current response time models are based on unidimensional IRT models and may not be able to account for a shared variance among testlet items.

The results strongly indicate the inaccuracy of parameter estimation that occurs when testlet effects (local dependencies) among items are ignored. The result of ignoring testlet effects is to greatly increase the amount of error in estimated parameters. The findings from the current study verify some previously identified effects. As with some previous studies (Bradlow, Wainer, & Wang, 1999; Im & Skorupski, 2014; Sireci, Thissen, & Wainer, 1991; Wainer & Wang, 2000; Yen, 1993), the presence of local dependence resulted in errors (bias and MSE) in parameter estimates. The HTRT model makes a unique contribution to the field of educational measurement by addressing the local dependency among testlet items using response time model. Overall, the HTRT model had very small measurement error on estimated parameters. The current results demonstrate that the Hierarchical Framework model had very good recovery of both the time intensity parameters, but fairly poor recovery of the item discrimination and time discrimination parameters. The examinee ability and speed parameters showed poor recovery, due not to bias but to dramatically increase random error. Finally, the HTRT model did indicate that parameters were estimated well across all ability level and the estimated test information was very close to the true value.

The response time model can support the test construction to balance time constraints. The response time item parameters can identify the amount of time each item demands from examinees. Even though the item difficulties can be similar, the location of items can have different consequence to the outcome. The items demanding a large amount of time at the end of assessment can have different impact to examinees than same items at the beginning. If the items

at the beginning of the assessment require a large amount of time, then examinees may face the speededness issue at the end. With the computerized adaptive testing, with respect to remaining assessment time, response time related information can support item selection procedure with similar item difficulties but require less response time. The additional source of information can potentially improve item selection and test construction to obtain increased precision of ability.

The response time models can furthermore provide feedback to examinees and test developers. Examinees can receive information to enhance their motivation and improve learning. Test developers can receive information to make modification and adjustment for valid and reliable assessment, and obtain meaningful results. The HTRT model can estimate item and person parameters just like the Hierarchical Framework model but also contains extra parameters to identify testlet effects. If the passages contain significant testlet effects, applying the model that cannot account for testlet effects can impact reliability and outcome.

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied psychological measurement, 21*, 1-23.

Baxter, B. (1941). An experimental analysis of the contributions of speed and level in an intelligence test. *Journal of Educational Psychology*, *32*, 285.

Bejar, I. (1985). *Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language* (Report No. ETS-RR-85-11). Princeton, NJ: Educational Testing Service.

Bergstrom, B., Gershon, R., & Lunz, M. E. (1994). *Computer-adaptive testing: Exploring examinee response time using hierarchical linear modeling*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 397-479). Reading, MA: MIT Press.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bontempo, B. D., & Julian, E. R. (1997). *Assessing speededness in variable-length computer adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Bridgeman, B. (2000). *Fairness in computer based testing: what we know and what we need to know.* (The GRE FAME Report). Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer adaptive test scores. *Journal of Educational Measurement, 41,* 137-148.

Bridges, K. R. (1985). Test-completion speed: Its relationship to performance on three course based objective examination, *Educational Psychological Measurement, 45*, 29-35.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics, 7*, 434-455.

Chang, S. (2006). *Computerized adaptive test item response times for correct and incorrect pretest and operational items: Testing fairness and test-taking strategies (*Unpublished doctoral dissertation). University of Nebraska, Lincoln.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Chen, C. T., & Wang, W. C. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement, 31*, 388-411.

Cronbach, L. J., & Warrington, W. G. (1951). Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika, 16*, 167-188.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement, 9,* 123-131.

Foos, P. W. (1989). Completion time and performance on multiple-choice and essay tests. *Bulletin of the Psychometric Society, 27,* 179-180.

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York:

Springer.

Fox, J. P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package CIRT. *Journal of Statistical Software, 20,* 1-14.

Gaviria, J. L. (2005). Increase in precision when estimating parameters in computer assisted testing using response time. *Quality and Quantity*, *39*, 45-69.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

Gelman, A. (1996). Inference and monitoring convergence. In W. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 131-143). London: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis (2nd ed.).* New York: Chapman & Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-472.

Gewke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculation of posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics* (pp. 169-193). Oxford: Oxford University Press.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Halkitis, P. N., Jones, J. P., & Pradhan, J. (1996). *Estimating testing time: The effects of item characteristics on response latency*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response

theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 253-261.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Hornke, L. (2000). Item response time in computerized adaptive testing. *Psicológica, 21,* 175-189.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods, 2*, 261-277.

Im, S. K., & Skorupski, W. P. (2014). *Parameter estimation error when ignoring testlet effects*. Graduate student research session at the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*, 331-358.

Keng, L., Ho, T., Chen, T. A., and Dodd, B. G. (2008). *A comparison of item and testlet selection procedures in computerized adaptive testing*. Paper presented at the Annual Meeting of the NCME, New York City, New York.

Kim, J. S., & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26,* 38-51.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8,* 147–154.

Lee, Y. H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*, 359-379.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied*

Psychological Measurement, 30, 3-21.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*, 353-373.

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution, 3*, 112-115.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New Jersey: Lawrence Erlbaum.

Luce, R. D. (1986). *Response times: their role in inferring elementary mental organization.* New York: Oxford University Press.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine, 28*, 3049-3067.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists.* New York: Springer.

Marianti, S., Fox, J. P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*, 426-451.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Ed.), *Handbook of modern item response theory* (pp. 101-121). New York: Springer.

Morison, E. J. (1960). On test variance and the dimensions of the measurement situation. *Education and Psychological Measurement, 20*, 231-250.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM algorithm.

Applied Psychological Measurement, 16, 159–176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17,* 351–363.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Ed.), *Handbook of modern item response theory* (pp. 153-164). New York: Springer.

Myers, C. T. (1952). The factorial composition and validity of differently speeded tests. *Psychometrika, 17,* 347-352.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31,* 200-219.

Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342-366.

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146-178.

Plake, B. (1999). *A new breed of CATS: Innovations in computerized adaptive testing*. Paper published by the University of Nebraska, Lincoln.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raftery, A.E., & Lewis, S.M. (1992). One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo. *Statistical Science, 7,* 493-497.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago:

University of Chicago Press.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement, 16,* 261-270.

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151-171). Amsterdam: North-Holland.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187-208). New York: Springer.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph, 17.*

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph, 18.*

Samejima, F. (1973). Homogeneous case of the continuous response level. *Psychometrika, 38,* 203-219.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39*, 111-121.

Samejima, F. (1983). *A general model for the homogeneous case of the continuous response* (ONR Research Report 83-3). Arlington, VA: Office of Naval Research, Personnel and Training Research Programs.

Samejima, F. (1997). Graded response model. In W.J. Van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Scheiblechner, H. (1979). Specific objective stochastic latency mechanisms. *Journal of Mathematical Psychology, 19,* 18-38.

Scheiblechner, H. (1985). Psychometric models for speed-test construction: The linear exponential model. In S. E. Embretson (Ed.), *Test design: Developments in psychology and education* (pp. 219–244). New York: Academic Press.

Schnipke, D. L., & Scrams, D. J. (1999). *Response-time feedback on computer administered tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Scrams, D. J., & Schnipke, D. L. (1997). *Making use of response times in standardized tests: Are accuracy and speed measuring the same thing?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Smith, R. (2000). *An exploratory analysis of item parameters and characteristics that influence item response time.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Spearman, C. (1927). *The abilities of man.* New York, NY: Macmillan.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1,* 81-87.

Sturtz, S., Ligges, U., & Gelman, A. (2010). R2OpenBUGS: A Package for Running OpenBUGS

from R. URL http:// cran.r-project.org/web/packages/R2OpenBUGS/vignettes/R2OpenBUGS.pdf

Suh, H. (2010). *A study of bayesian estimation and comparison of response time models in item response theory* (Unpublished doctoral dissertation). University of Kansas, Lawrence, KS.

Swygert, K. A. (1998). *An examination of item response times on the GRE-CAT. (*Unpublished doctoral dissertation). University of North Carolina, Chapel Hill.

Tatsuoka, K. K., & Tatsuoka, M. M. (1980). A model for incorporating response-time data in scoring achievement tests. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 236-256). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247-260.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological methods, 6*, 181-195.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31,* 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 73,* 287-308.

van der Linden,W. J. (2009). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement, 33*, 25-41.

van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement, 48*, 44-60.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement, 44*, 117–130.

van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W.J. Van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1-28). New York: Springer.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement, 23*, 195-210.

van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika, 68*, 251-265.

van der Linden, W. J., & Xiong, X. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics, 39*, 418-438.

Verhelst, N. D., Verstraalen, H. H. F. M., & Jansen, M. G. (1997). A logistic model for time limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169-185). New York: Springer.

Wainer, H. (1994). *A testlet-based examination of the LSAT* (Statistical Report 93-03). Newtown, PA: Law School Admission Council.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas

(Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Boston, MA: Kluwer Academic Publishers.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-202.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, *27*, 1-14.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*, 22-29.

Wainer, H. & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203-220.

Wang, X., Baldwin, S., Wainer, H., Bradlow, E. T., Reeve, B. B., Smith, A. W., Bellizzi, K. M., and Baumgartner, K. B. (2010). Using testlet response theory to analyze data from a survey of attitude change among breast cancer survivors. *Statistics in Medicine*, *29*, 2028-2044.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*, 109-128.

Wang, T. & Hanson, B. A (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29,* 323-339.

Wang, W., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126-149.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43,* 19-38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163-183.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40,* 307-330.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Ed.), *Applications of latent trait and latent class models in the social sciences* (pp. 89-98). New York: Waxman.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–213.

Zenisky, A. L., & Baldwin, P. (2006). *Using response time data in test development and validation: Research with beginning computer users.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Zenisky, A.L., Hambleton, R. K., & Sireci, S.G. (2002). Identification and evaluation of local item dependencies in the Medical College Admission Test. *Journal of Educational Measurement, 39,* 291-309.

Appendix A: Scatter plots of comparable parameters

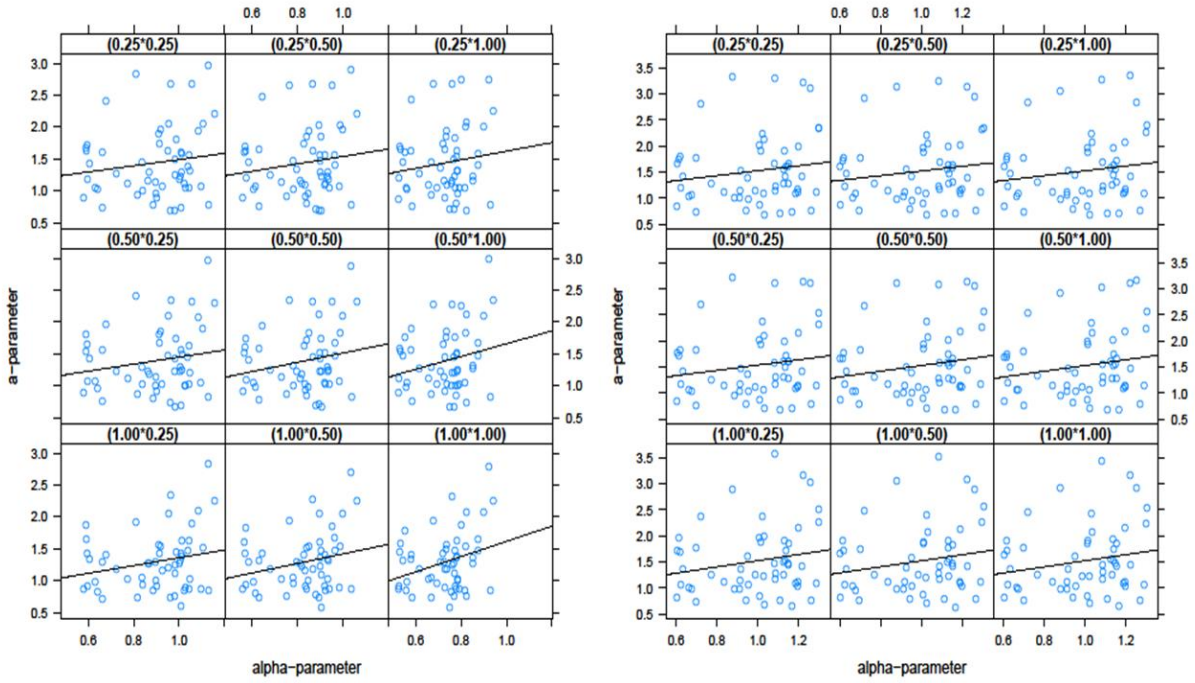

Figure A1. *Scatter plots using parameter estimates between a-parameter and alpha-parameter of the HTRT (right) and the Hierarchical Framework (left) for all nine conditions in grade 3.*

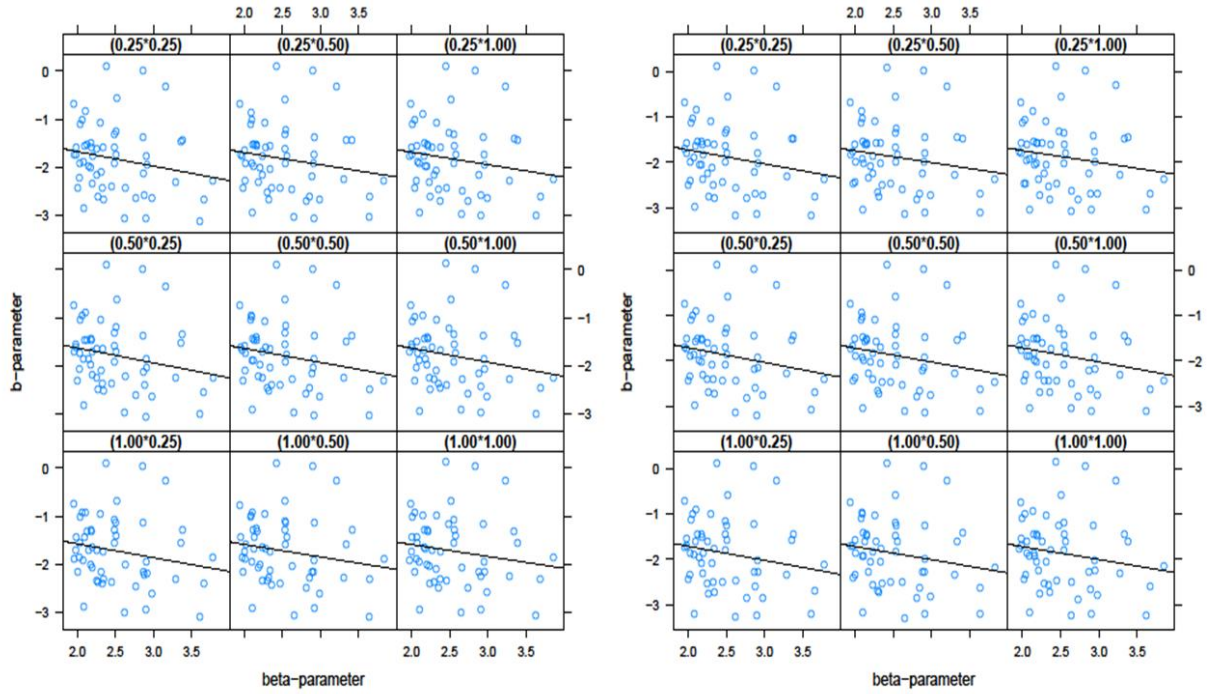

Figure A2. *Scatter plots using parameter estimates between b-parameter and beta-parameter of the HTRT (right) and the Hierarchical Framework (left) for all nine conditions in grade 3.*

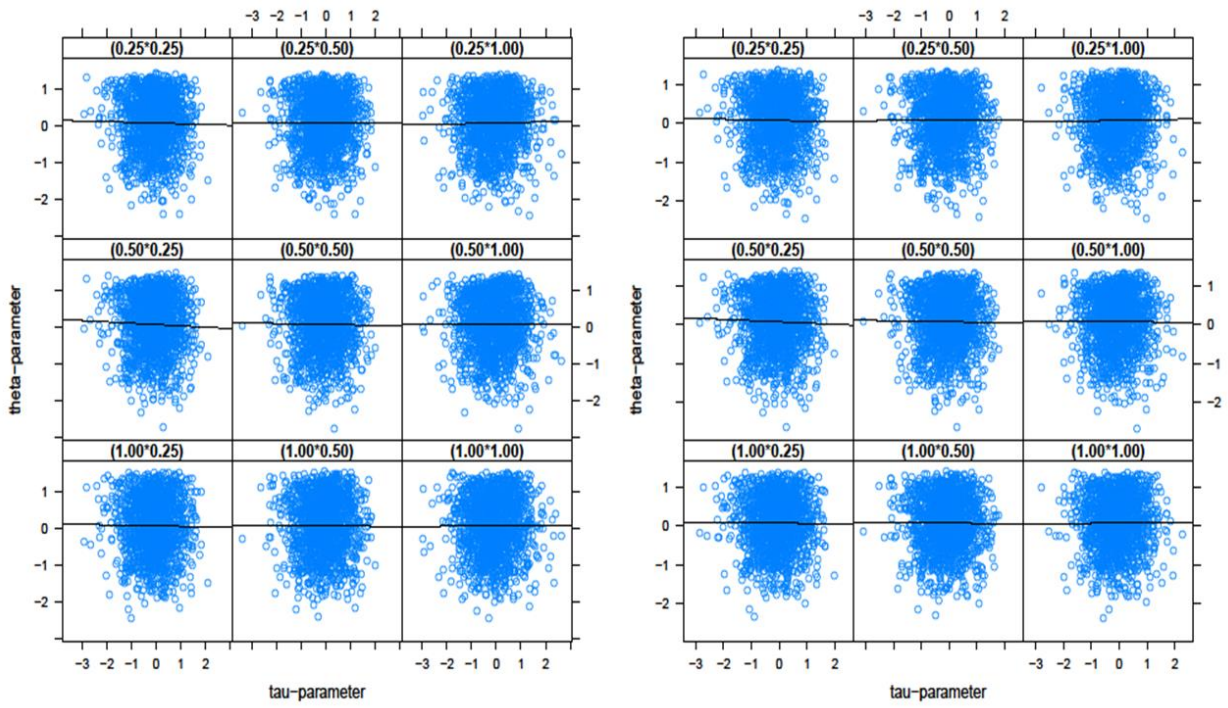

Figure A3. *Scatter plots using parameter estimates between theta-parameter and tau-parameter of the HTRT (right) and the Hierarchical Framework (left) for all nine conditions in grade 3.*

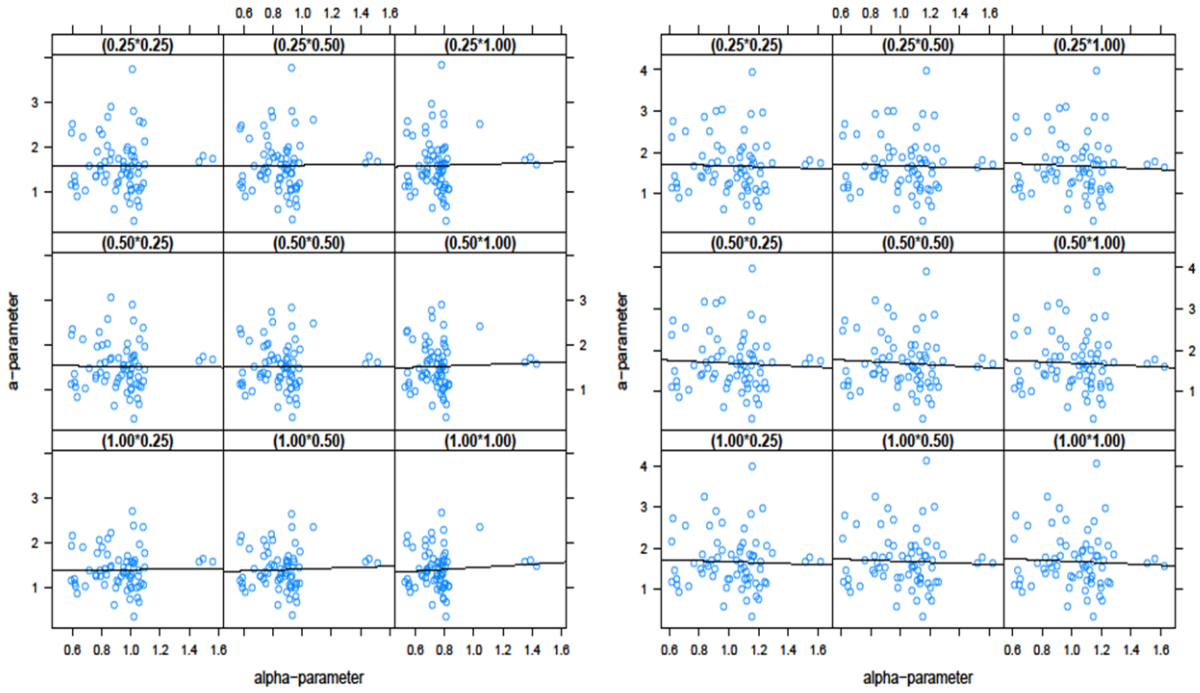

Figure A4. *Scatter plots using parameter estimates between a-parameter and alpha-parameter of the HTRT (right) and the Hierarchical Framework (left) for all nine conditions in grade 4.*

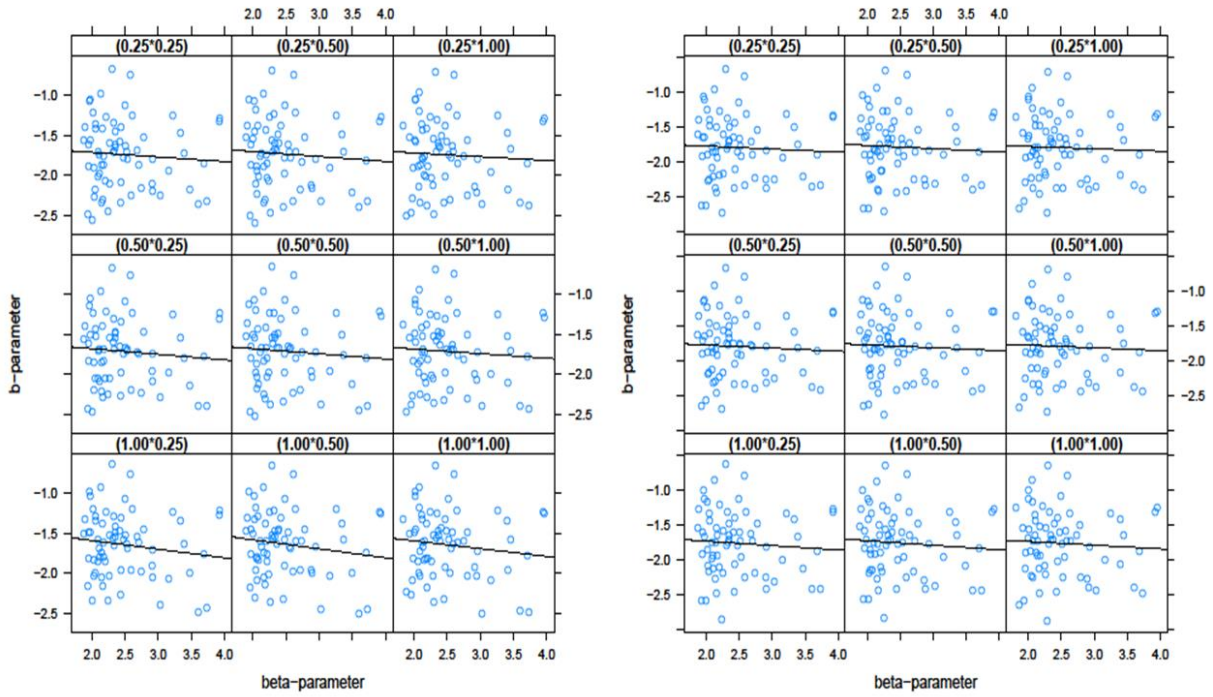

Figure A5. *Scatter plots using parameter estimates between b-parameter and beta-parameter of the HTRT (right) and the Hierarchical Framework (left) for all nine conditions in grade 4.*

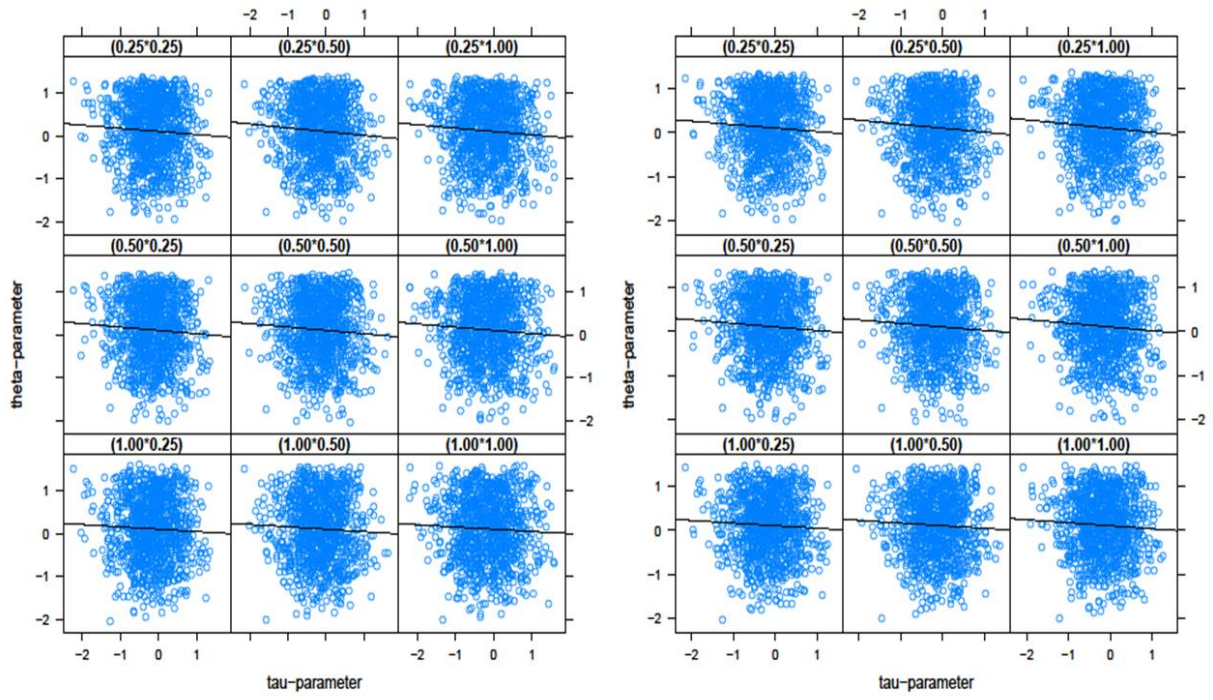

Figure A6. *Scatter plots using parameter estimates between theta-parameter and tau-parameter of the HTRT (right) and the Hierarchical Framework (left) for all nine conditions in grade 4.*

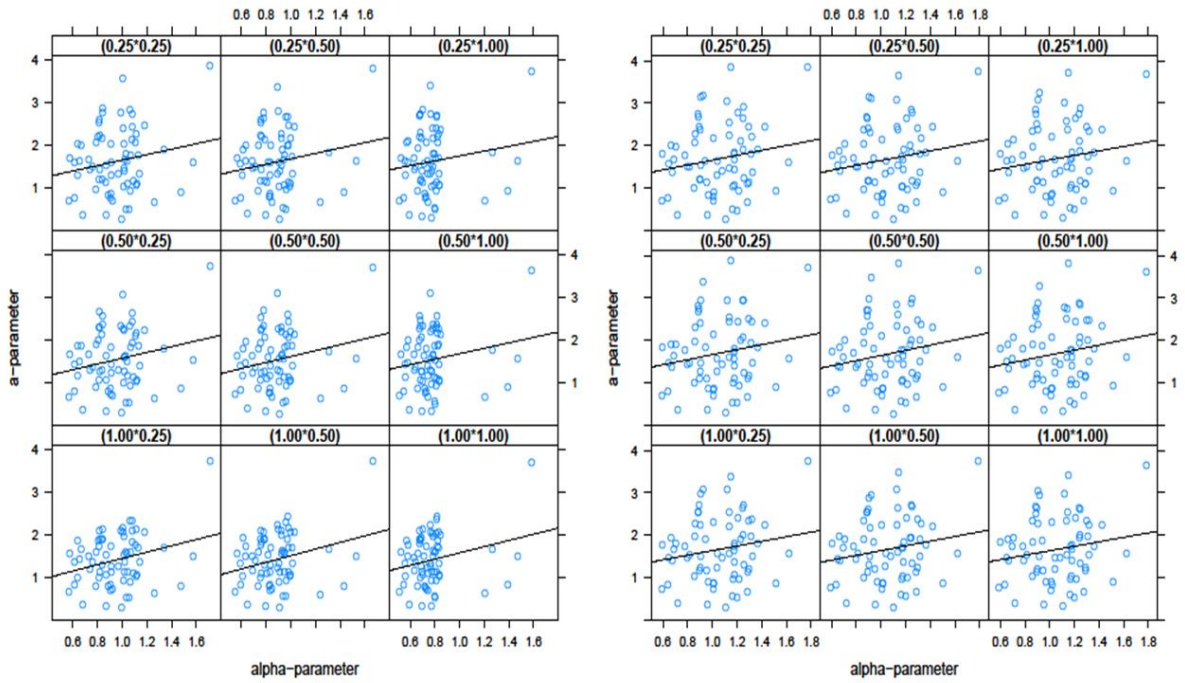

Figure A7. *Scatter plots using parameter estimates between a-parameter and alpha-parameter of the HTRT (right) and the Hierarchical Framework (left) for all nine conditions in grade 5.*

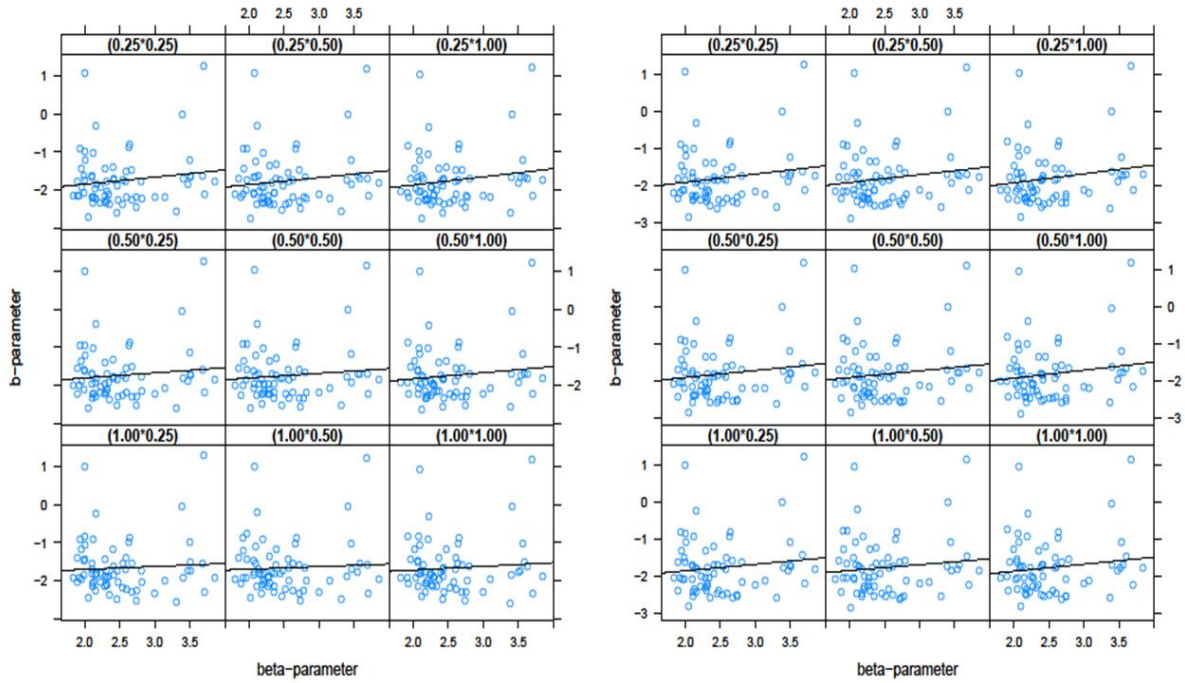

Figure A8. *Scatter plots using parameter estimates between b-parameter and beta-parameter of the HTRT (right) and the Hierarchical Framework (left) for all nine conditions in grade 5.*

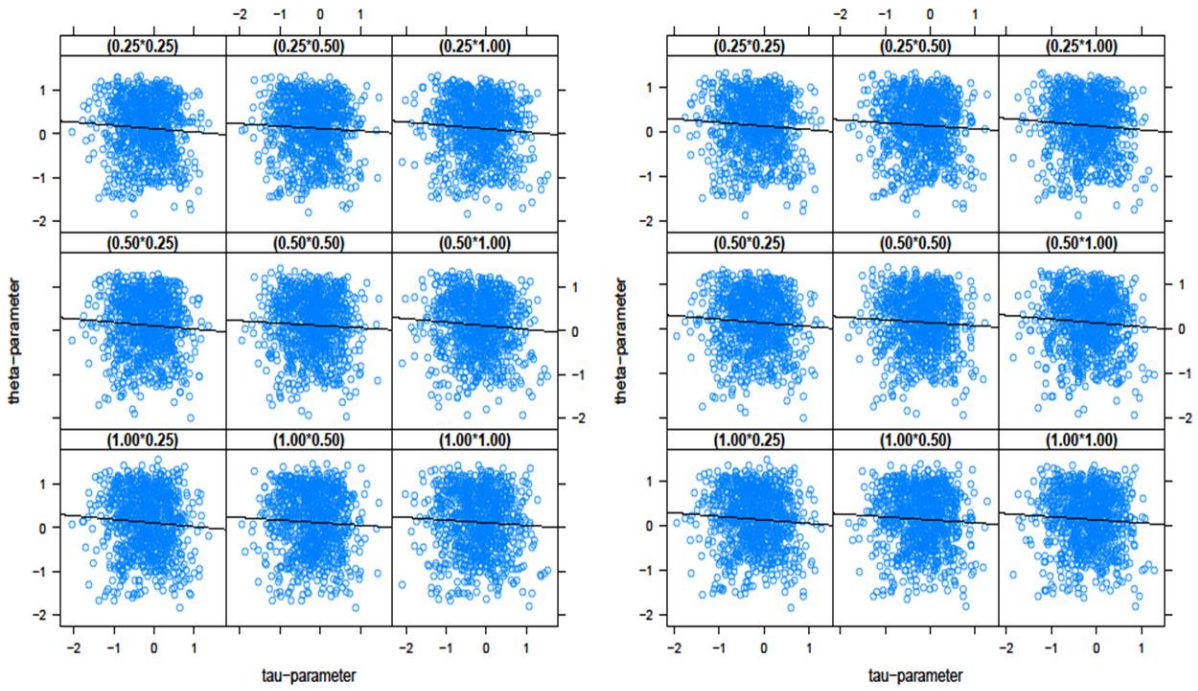

Figure A9. *Scatter plots using parameter estimates between theta-parameter and tau-parameter of the HTRT (right) and the Hierarchical Framework (left) for all nine conditions in grade 5.*

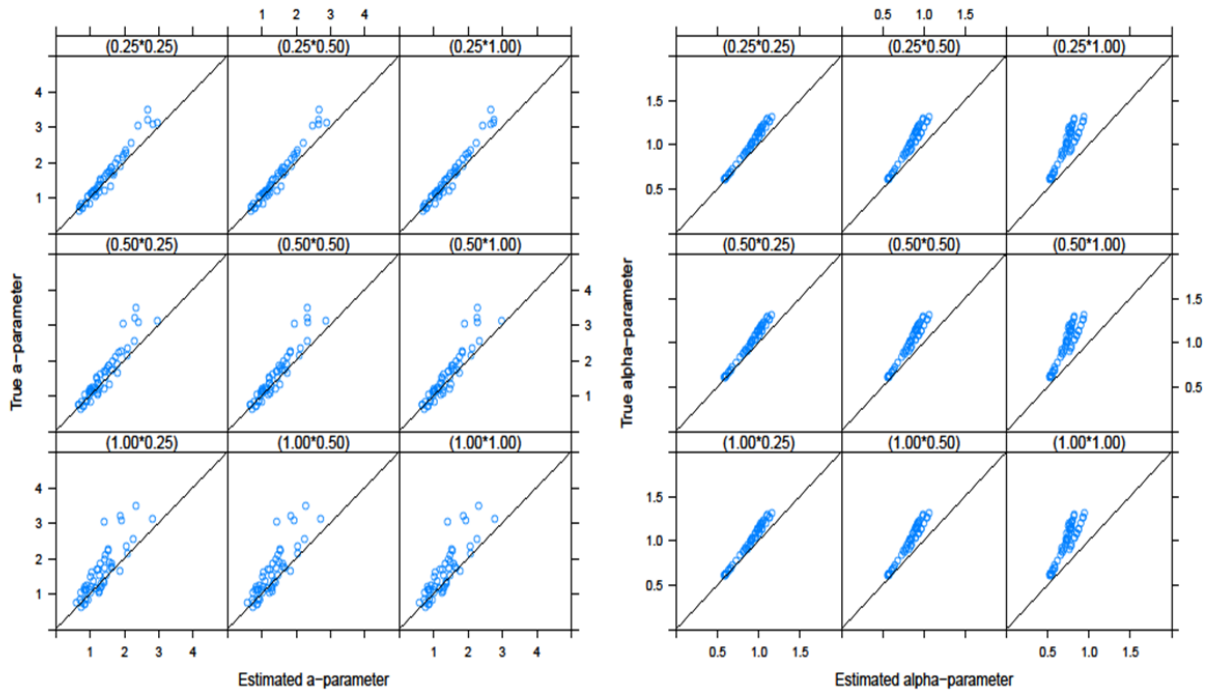

Figure A10. *The Hierarchical Framework model's scatter plots between true parameters and estimated a- (left) and alpha- (right) parameters for all nine conditions in grade 3.*

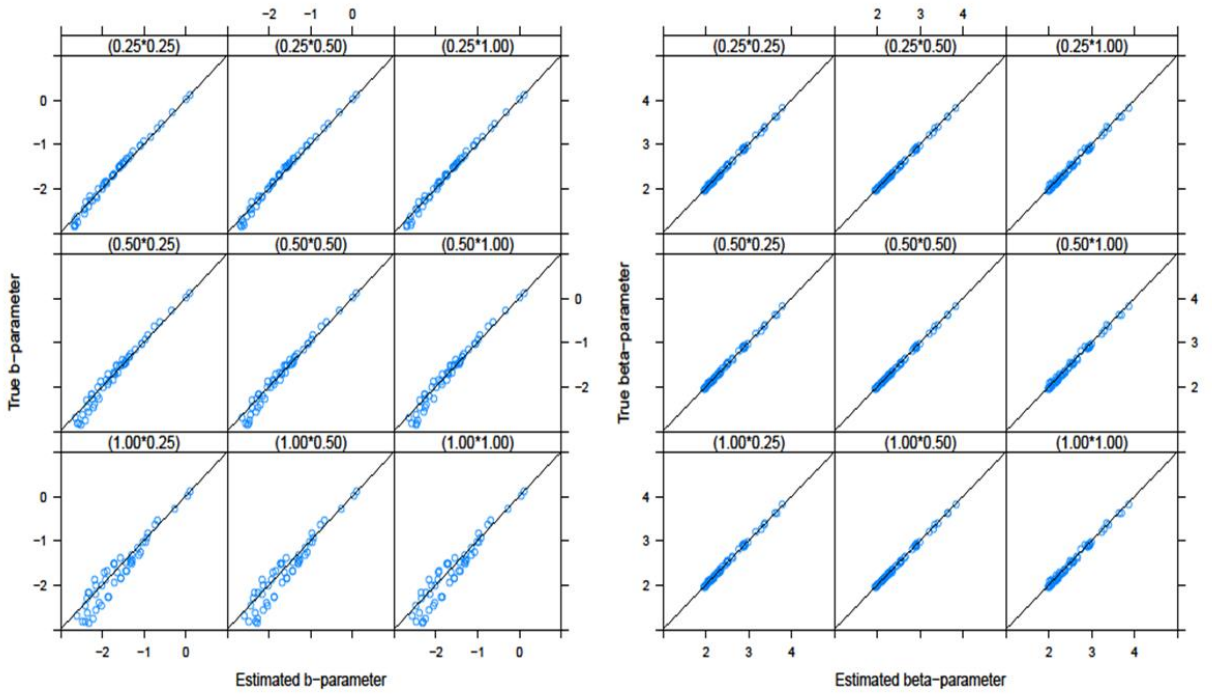

Figure A11. *The Hierarchical Framework model's* scatter plots between true parameters and estimated b- (left) and beta- (right) parameters for all nine conditions in grade 3.

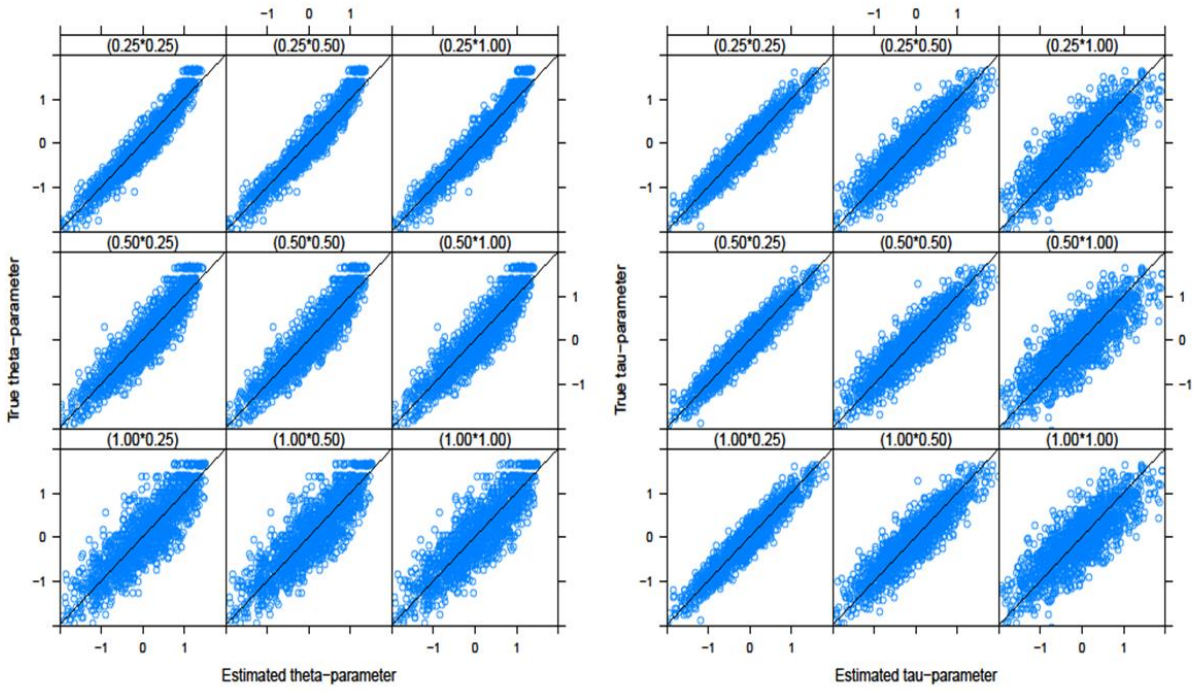

Figure A12. *The Hierarchical Framework model's* scatter plots between true parameters and estimated theta- (left) and tau- (right) parameters for all nine conditions in grade 3.

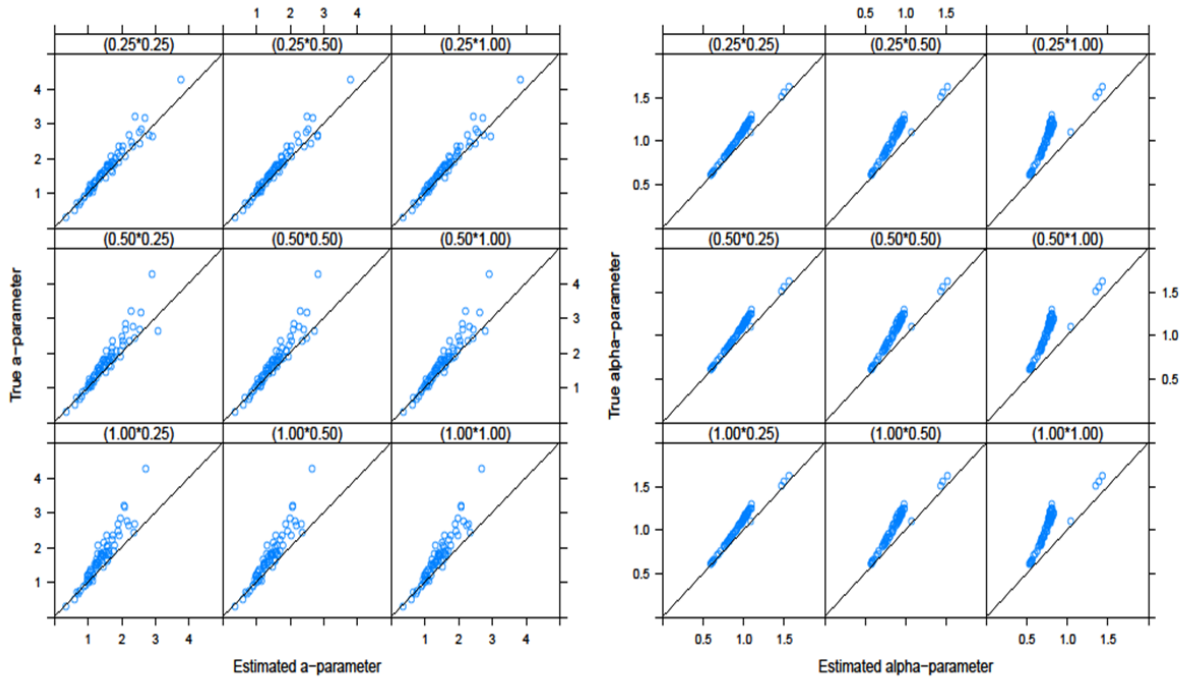

Figure A13. *The Hierarchical Framework model's* scatter plots between true parameters and estimated a- (left) and alpha- (right) parameters for all nine conditions in grade 4.

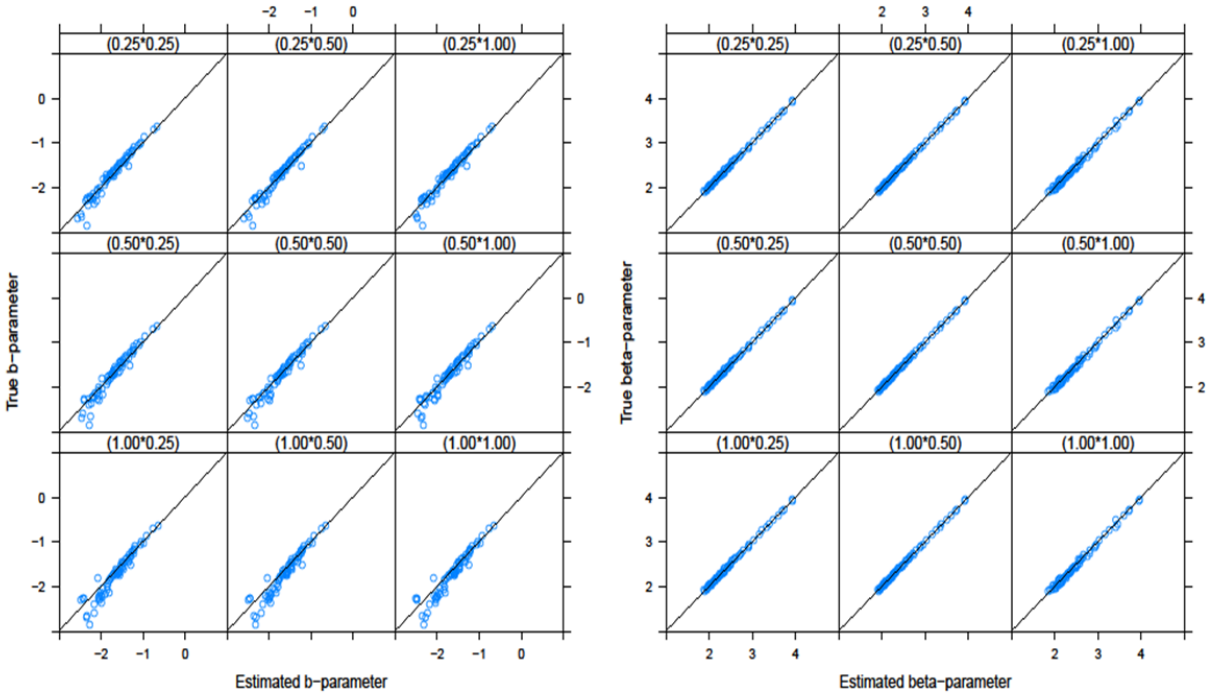

Figure A14. *The Hierarchical Framework model's* scatter plots between true parameters and estimated b- (left) and beta- (right) parameters for all nine conditions in grade 4.

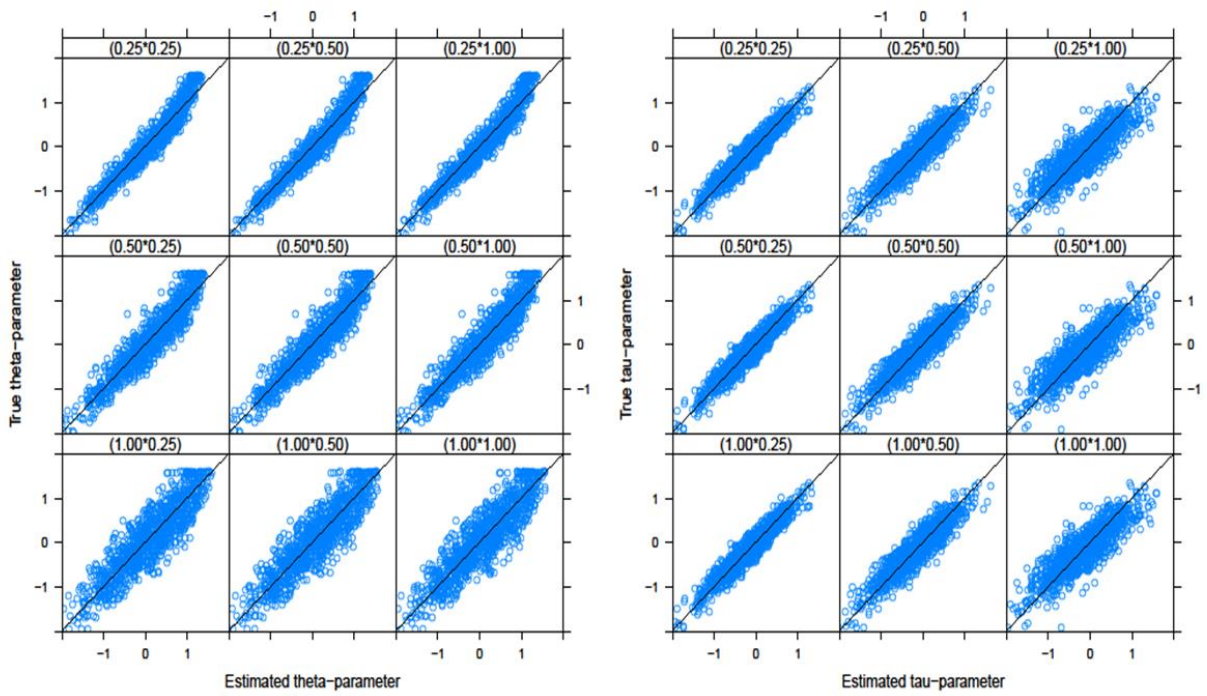

Figure A15. *The Hierarchical Framework model's* scatter plots between true parameters and estimated theta- (left) and tau- (right) parameters for all nine conditions in grade 4.

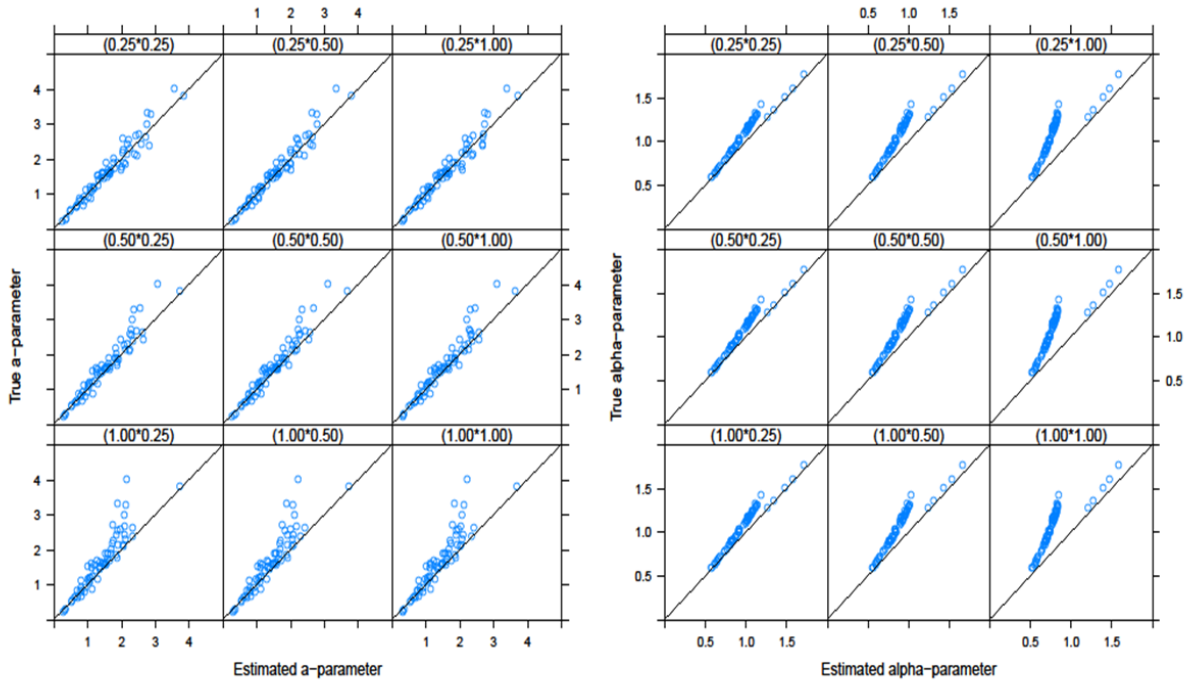

Figure A16. *The Hierarchical Framework model's* scatter plots between true parameters and estimated a- (left) and alpha- (right) parameters for all nine conditions in grade 5.

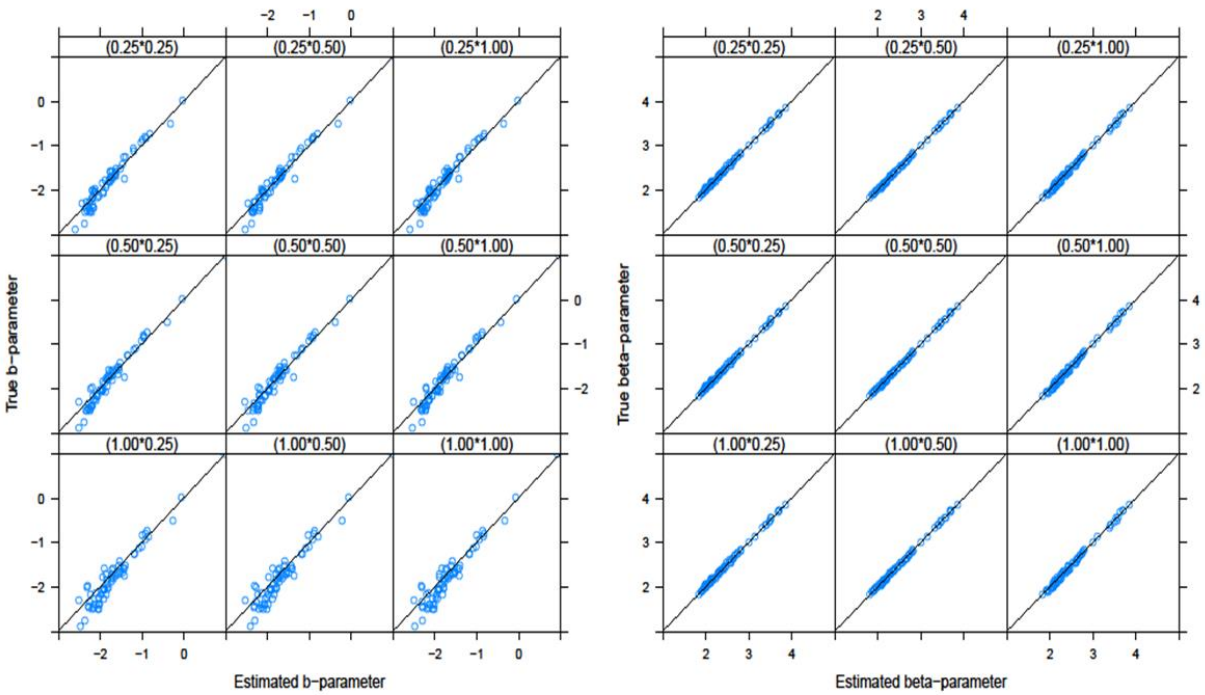

Figure A17. *The Hierarchical Framework model's* scatter plots between true parameters and estimated b- (left) and beta- (right) parameters for all nine conditions in grade 5.

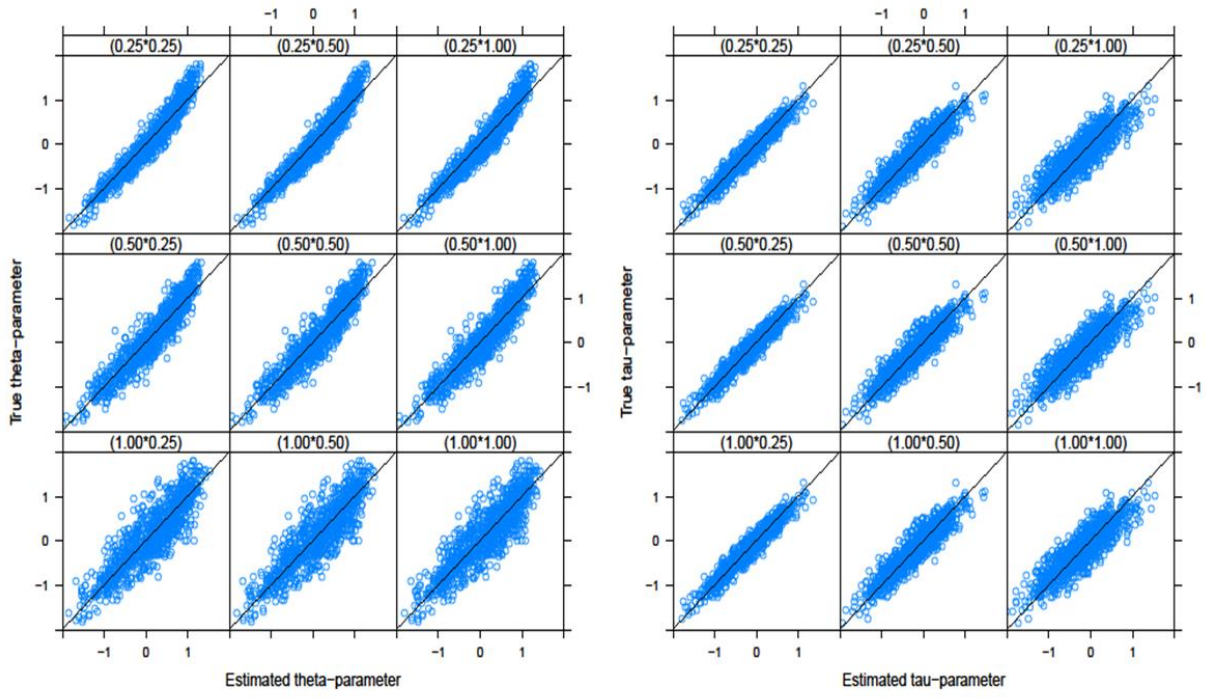

Figure A18. *The Hierarchical Framework model's* scatter plots between true parameters and estimated theta- (left) and tau- (right) parameters for all nine conditions in grade 5.

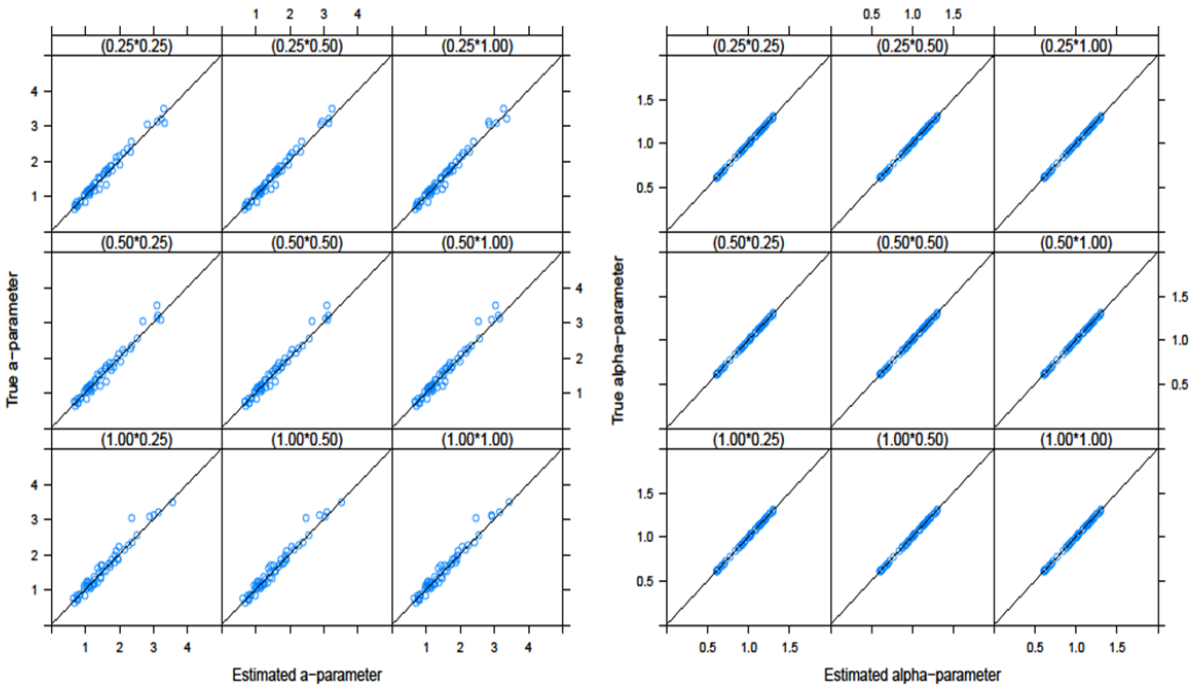

Figure A19. *The HTRT model's* scatter plots between true parameters and estimated a- (left) and alpha- (right) parameters for all nine conditions in grade 3.

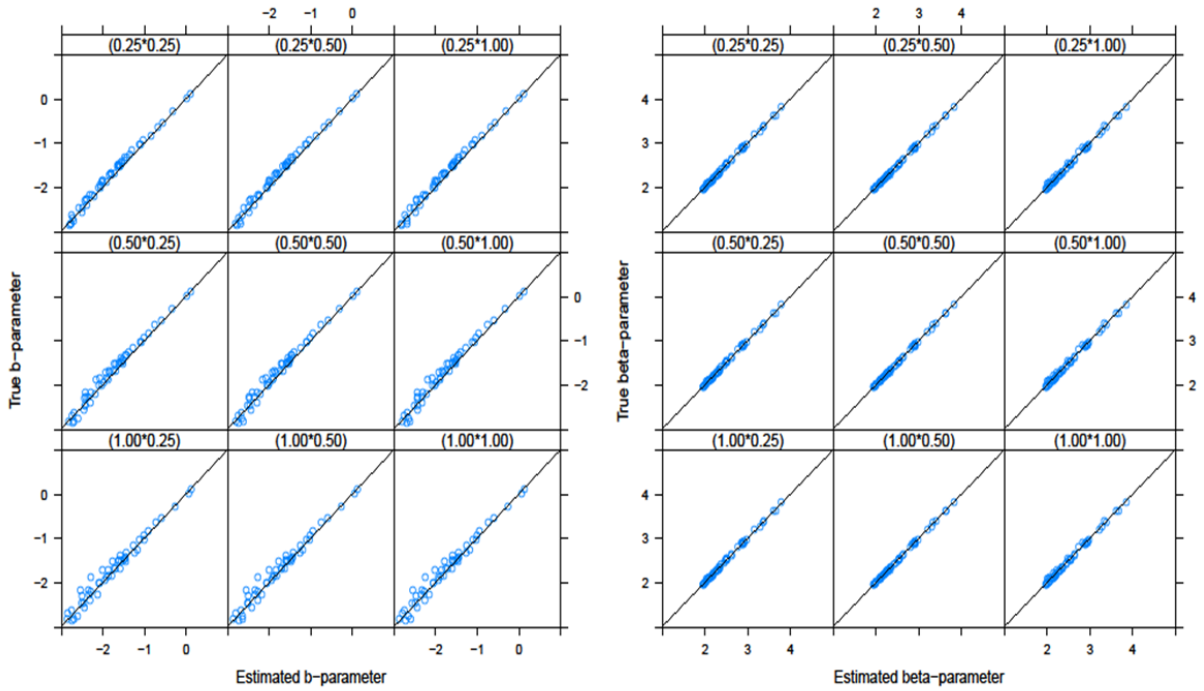

Figure A20. *The HTRT model's* scatter plots between true parameters and estimated b- (left) and beta- (right) parameters for all nine conditions in grade 3.

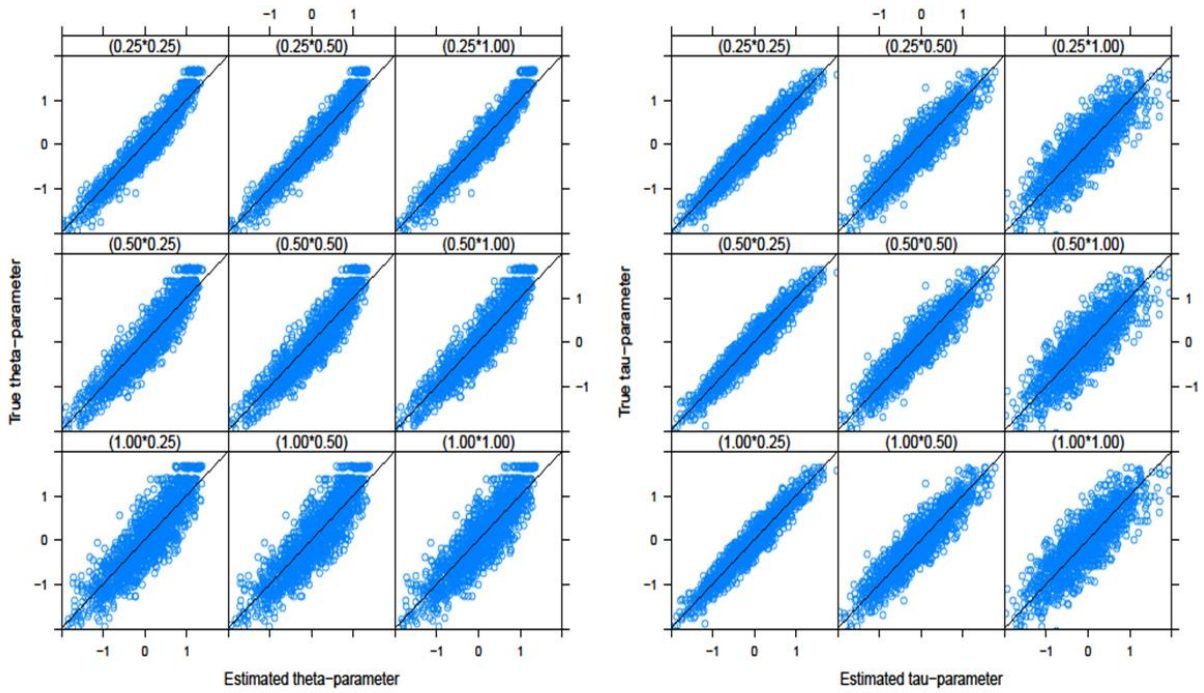

Figure A21. *The HTRT model's scatter plots between true parameters and estimated theta- (left) and tau- (right) parameters for all nine conditions in grade 3.*

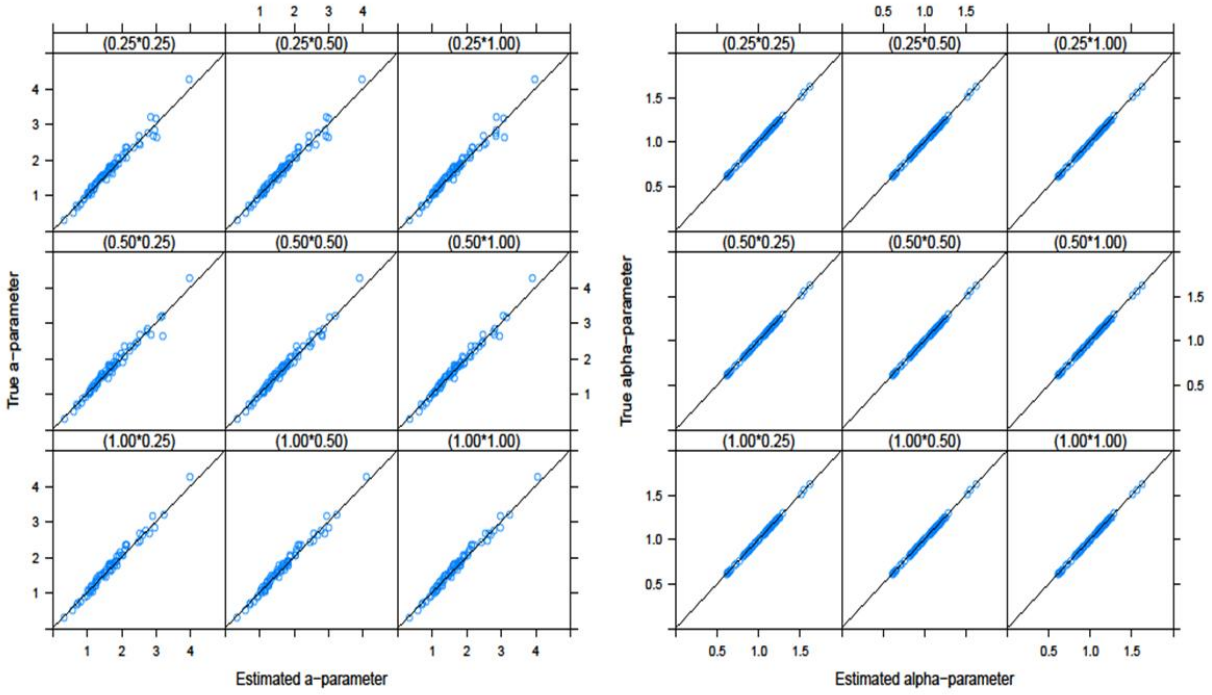

Figure A22. *The HTRT model's scatter plots between true parameters and estimated a- (left) and alpha- (right) parameters for all nine conditions in grade 4.*

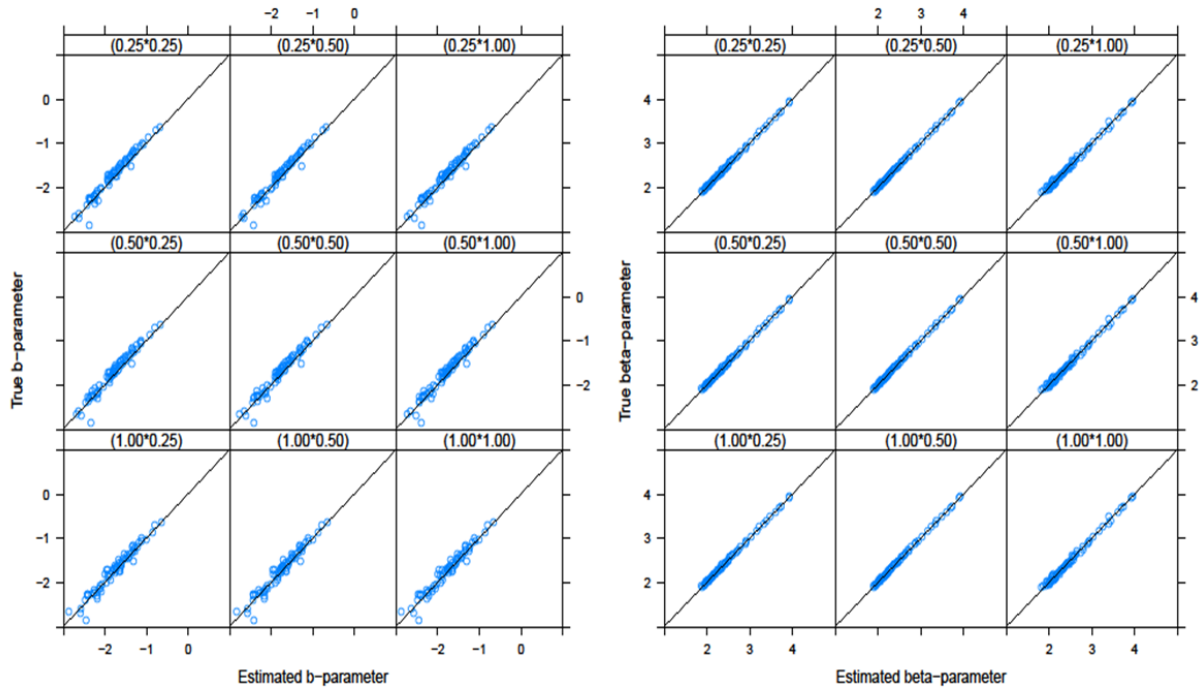

Figure A23. *The HTRT model's* scatter plots between true parameters and estimated b- *(left) and beta- (right) parameters for all nine conditions in grade 4.*

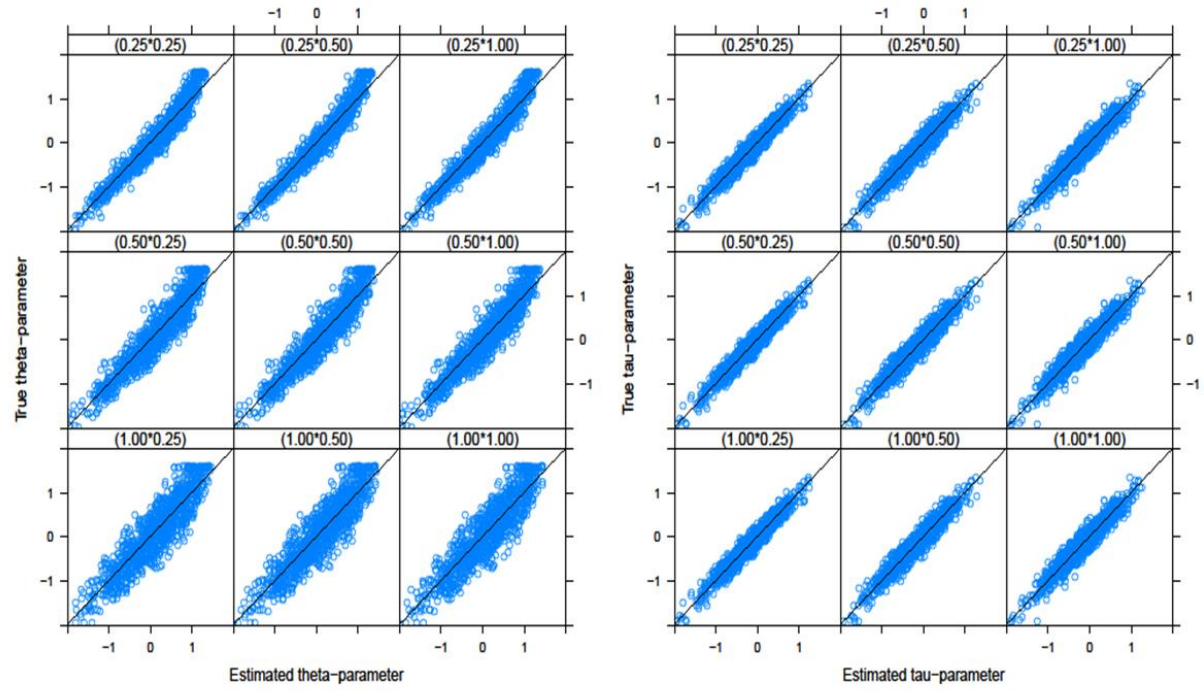

Figure A24. *The HTRT model's* scatter plots between true parameters and estimated theta- *(left) and tau- (right) parameters for all nine conditions in grade 4.*

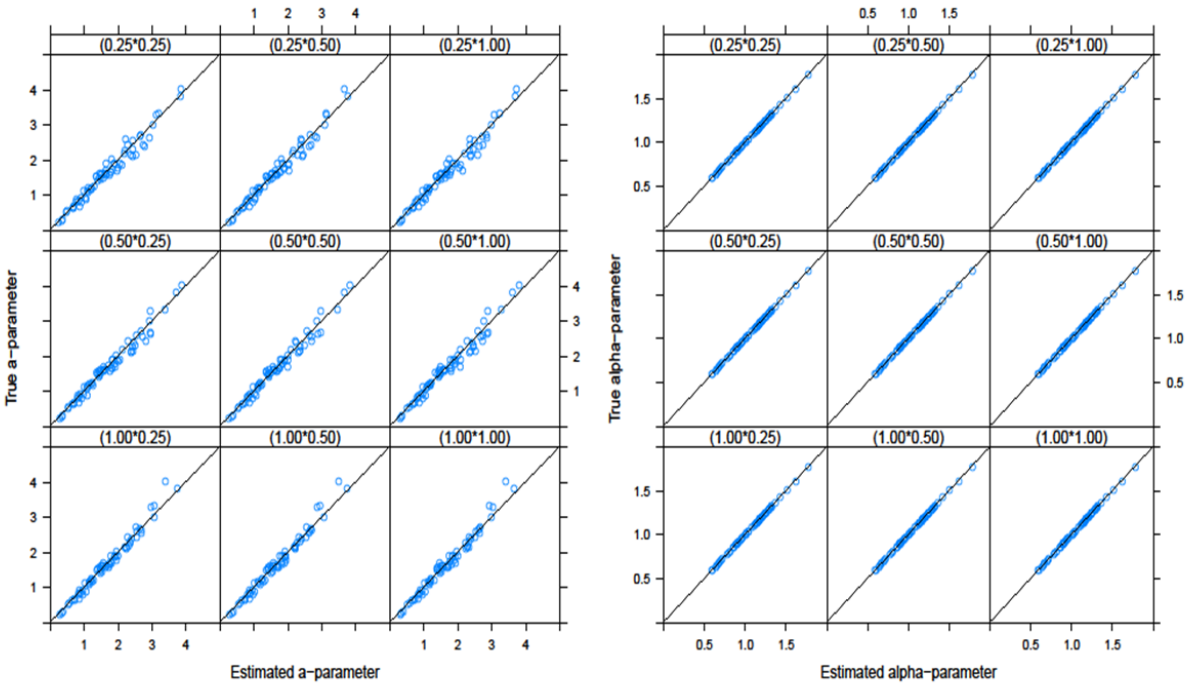

Figure A25. *The HTRT model's* scatter plots between true parameters and estimated a- *(left) and alpha- (right) parameters for all nine conditions in grade 5.*

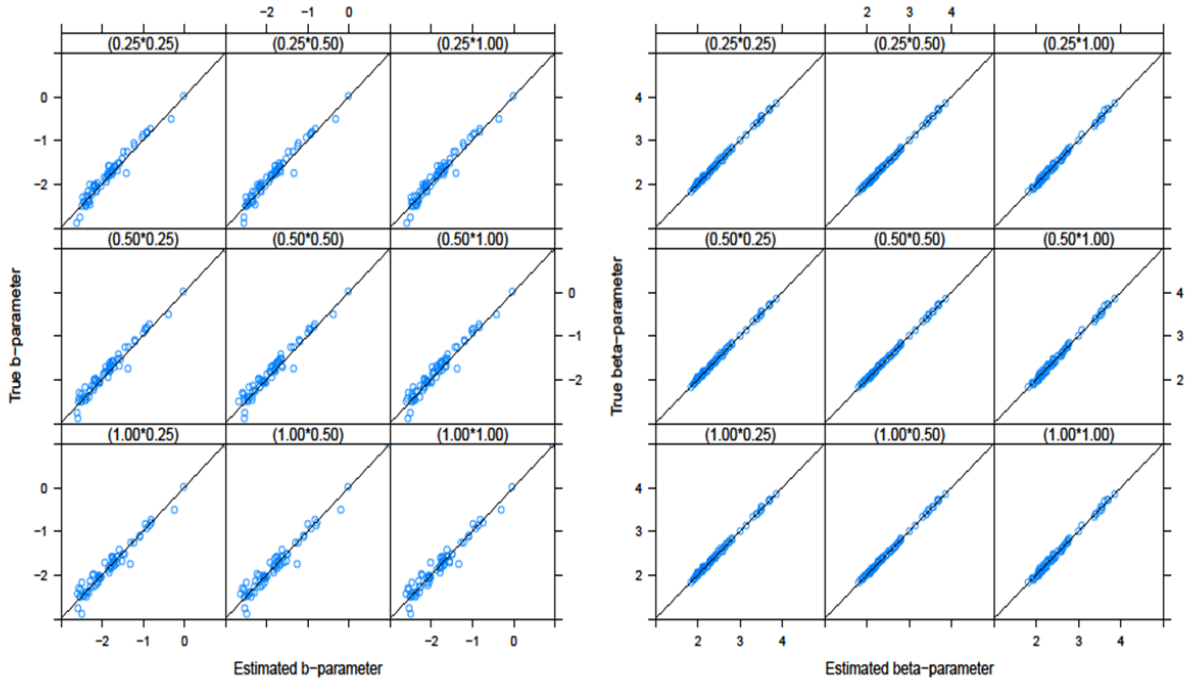

Figure A26. *The HTRT model's* scatter plots between true parameters and estimated b- *(left) and beta- (right) parameters for all nine conditions in grade 5.*

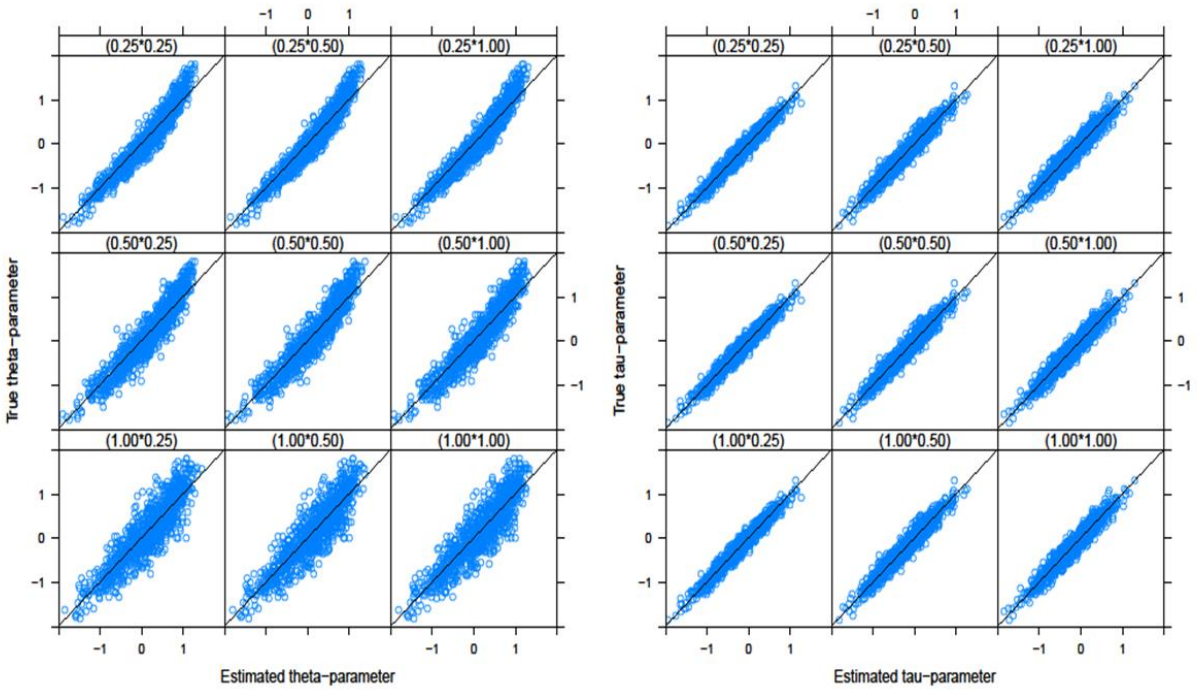

Figure A27. *The HTRT model's* scatter plots between true parameters and estimated theta- *(left) and tau- (right) parameters for all nine conditions in grade 5.*